

AD-A097 896

TEXAS UNIV AT AUSTIN LINGUISTICS RESEARCH CENTER

F/G 5/7

THE METAL SYSTEM. VOLUME I AND VOLUME II. APPENDICES. (U)

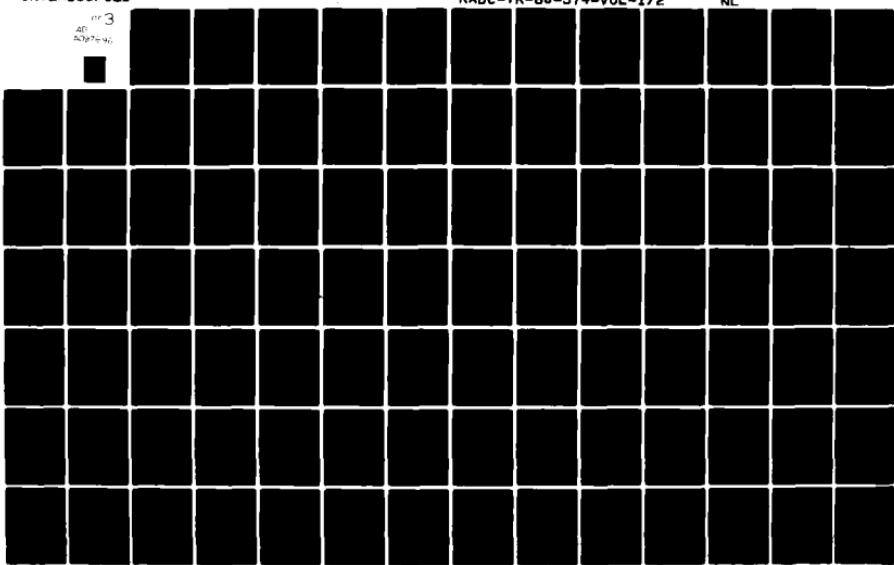
JAN 81 W P LEHMANN, W S BENNETT, J SLOCUM F30602-78-C-0132

UNCLASSIFIED

RADC-TR-80-374-VOL-1/2

NL

m3  
40  
20287-40



A.583.22 [Satz-Buendel-Messung bei Fernbetrieb]

A.583.22

[Sollen alle Buendel] E1  
|der VST gemessen werden?] N

J

[Soll die] E2  
|Ausgabe auf MB  
erfolgen?] N

MBG 01

J

[Magnetband  
einhaengen]

DSS 02

DSS 03

[Satz-Buendel-Messung  
mit  
{BA = BINAER}  
starten]

[Satz-Buendel-Messung  
mit  
{BA = ASCII}  
starten]

[Ende]

I-----I  
I A.583.22.E1 I [Sollen alle Buendel der VST gemessen werden?] I-----I

[Wenn "Ja" muss die Ausgabe auf Magnetband erfolgen.] [Weiter mit S01.]

["Nein" bedeutet, dass nur ausgewahlte Buendel der VST gemessen werden.] [Weiter mit E2.]

I-----I  
I A.583.22.E2 I [Soll die Ausgabe auf Magnetband erfolgen?] I-----I

[Bei Messungen ausgewahlter Buendel kann die Ausgabe auf Magnetband oder Drucker der TD-EWS erfolgen.]

**Figure 6-2: Sample Page of Prepared German Telecommunications Text**

A.583.22 [0763#####]

A.583.22

[0764#####] E1  
#####]

J

[0765#####] E2  
#####  
#####]

N

MBG 01

J

[0766#####]  
#####]

DSS 02

DSS 03

[0767#####  
###  
#####]  
#####]

[0768#####  
###  
#####]  
#####]

[0769###]

I-----I  
I A.583.22.E1 I [0770#####]

I-----I  
[0771#####] [0772#####]  
#####]

[0773#####]  
#####] [0774#####]

I-----I  
I A.583.22.E2 I [0775#####]

[0776#####]  
#####]

Figure 6-3: Text Format Remaining in Document After Extraction

(0763 Satz-Buendel-Messung bei Fernbetrieb)  
(0764 Sollen alle Buendel der VST gemessen werden?)  
(0765 Soll die Ausgabe auf MR erfolgen?)  
(0766 Magnetband einhaengen)  
(0767 Satz-Buendel-Messung mit {BA = BINAER} starten)  
(0768 Satz-Buendel-Messung mit {BA = ASCII} starten)  
(0769 Ende)  
(0770 Sollen alle Buendel der VST gemessen werden?)  
(0771 Wenn "Ja" muss die Ausgabe auf Magnetband erfolgen.)  
(0772 Weiter mit S01.)  
(0773 "Nein" bedeutet, dass nur ausgewahlte Buendel der VST gemessen werden.)  
(0774 Weiter mit E2.)  
(0775 Soll die Ausgabe auf Magnetband erfolgen?)  
(0776 Bei Messungen ausgewaehlerter Buendel kann die Ausgabe auf Magnetband oder Drucker der TD-EWS erfolgen.)

Figure 6-4: German Items to Be Translated as Extracted From Sample Page of Prepared German Telecommunications Text

(0763 peripheral circuit trunk group measurement in the case of  
remote operation)  
(0764 should all trunk groups of the VST be measured ?)  
(0765 should the output occur on MB ?)  
(0766 hang up magnetic tape)  
(0767 start peripheral circuit trunk group measurement with BA = BINAER)  
(0768 start peripheral circuit trunk group measurement with BA = ASCII)  
(0769 end)  
(0770 should all trunk groups of the VST be measured ?)  
(0771 if "Ja" the output on magnetic tape must occur .)  
(0772 further with S01 .)  
(0773 mean "Nein" , that only selected trunk groups of the VST are measured .)  
(0774 further with E2 .)  
(0775 should the output occur on magnetic tape ?)

Figure 6-5: English Items as Translated Before Reconstitution Into Text

## A.583.22 Peripheral circuit trunk group measurement in the case of remote operation

A.583.22

Should all trunk E1 groups of the VST be measured? N

J

Should E2  
the output  
occur on MB?

N

N

MBG 01

J

Hang up  
magnetic tape

DSS 02

DSS 03

Start peripheral  
circuit  
trunk group  
measurement with BA = BINAER

Start peripheral  
circuit  
trunk group  
measurement with BA = ASCII

End

I-----

I A.583.22.E1 I Should all trunk groups of the VST be measured?

I-----]

If "Ja" the output on magnetic tape must occur.  
Further with S01.

Mean "Nein", that only selected trunk groups of the VST are measured. Further with E2.

[-----]

I A.583.22.E2 I Should the output occur on magnetic tape?

-----

Figure 6-6: Reconstituted English Text Before Human Revision

A.583.22 Peripheral circuit trunk group measurement in the case of remote operation

A.583.22

Should all trunk groups El  
of the VST be measured? N

J

Should E2  
the output  
occur on MB? N

MBG 01

J

Hang up  
magnetic tape

DSS 02

DSS 03

Start peripheral  
circuit trunk group  
measurement with  
BA = BINAER

Start peripheral  
circuit trunk group  
measurement with  
BA = ASCII

End

I-----I  
I A.583.22.E1 I Should all trunk groups of the VST be measured?  
I-----I

If "Ja" the output must occur on magnetic tape.  
Continue with S01.

"Nein" means that only selected trunk groups of the VST are measured. Continue with E2.

I-----I  
I A.583.22.E2 I Should the output occur on magnetic tape?  
I-----I

The output may occur on magnetic tape or on the printer of the TD-EWS in the case of measurements of selected trunk groups.

Figure 6-7: Sample Page of English Text After Human Revision

## 7. METAL Applied to a 50-Page Text

In May 1980 LRC conducted a feasibility study regarding the question: Can METAL be applied to production MT on a cost-effective basis? METAL was applied to the translation of a 50-page German telecommunications text provided by the Siemens Corporation. We first summarize the quantitative results, then turn to an assessment of the quality and a discussion of cost-effectiveness.

### 7.1. Quantitative Results

Tables 7.1-7.4 summarize the machine performance statistics gathered during the experiment. Each sentence was tabulated according to its length and the type of success or failure experienced in attempting to translate it. For the analysis phase, measurements included the number of words in each sentence, the number of such sentences, the number of interpretations for each sentence, the number of grammatical phrases instantiated, the number of such phrases rejected on "semantic" grounds, the pure CPU time (exclusive of storage management time) expended, the number of LISP cells (machine words) allocated, and the fair CPU time expended. Similarly for the transfer/generation phase, the pure CPU time, storage demands, and fair CPU time were measured.

The fair measure reported here includes the pure CPU time plus a pro-rated portion of the time consumed by the LISP storage manager, as determined by the number of cells (memory words) used per sentence. In a system with limited address space, such as our DEC KI-10, the large time constants involved in storage management will, due to its relatively high incidence, significantly increase the actual processing time for a given task. We feel it only fair to include such overhead in our CPU time measurements, because one is billed for storage management time as well as pure processing time. However, on a machine with a larger address space the number of invocations of the storage manager would decrease dramatically and, all other things being equal, result in a significant decrease in fair CPU times. As an example of the variability in this figure, improvements in METAL's storage efficiency regarding lexical entries resulted in there being more than twice as many cells available for analysis purposes. This produced a 27% reduction in storage management time per cell unit as a consequence of the reduced incidence of storage management. The pure time figures, then, may be regarded as lower limits given the current METAL system and a KI-10-equivalent CPU having a larger address space.

Tables 7.1-7.2 summarize the data for the items successfully analyzed. There are two categories: (1) 213 1-word items without inflection, which were found in the lexicon and translated directly (see Table 7.1), and (2) 780 items for which recourse to parsing was necessary, either because a 1-word item was inflected (or unknown) or because the item was a phrase or sentence (see Table 7.2). The longest sentence analyzed contained 34 words. (There were only three longer sentences in the text.) The fair translation times in Table 7.2 can be approximated by a piece-wise linear function: for sentences up to 7 words the slope is about 3 seconds/word, after which the slope increases to about 5 seconds/word. Since 62% of these sentences are less than 7 words long, the overall average slope is 3.66 seconds/word. There are four anomalies in these figures: sentences of 15, 16, 17, and 34 words. A study of the raw data

indicates that these are in each case due to a particular sentence which is highly ambiguous according to our current grammar, and which therefore requires abnormally large amounts of space and CPU time. The most important point to observe is that the times are indeed linear, not polynomial or exponential. This means that normal improvements can be expected to have predictable and measurable benefits in terms of reduced processing requirements. Another important point, discussed later, is that these times, coupled with the measured translation quality, imply a substantial advance in the state of the art in MT.

Table 7.3 summarizes the data for the 91 sentences not successfully analyzed due to any of a variety of reasons, but which were translated phrase-by-phrase, after "phrasal analysis" by the special procedure USER\_POSTPROCESSOR (see Chapter V). A graph of these data also demonstrates linear behavior, with an average slope of 3.71 seconds/word. There is one anomaly: a single 28-word sentence required much more than the expected amount of processing.

Table 7.4 summarizes the data for 19 sentences for which the translation attempt resulted in a complete failure of the program in that METAL produced no translation. These occurred in the analysis phase, with one exception occurring in the transfer phase. Essentially the only meaningful data here are the number of sentences of each length for which this occurred, and the total number of such instances.

## 7.2. Quality Assessment

The 213 successful 1-word direct translations, plus the 780 successful analyses with translations, constitute 90% of the sentences in the 50-page text. The 91 analysis failures with phrasal translations constitute 8.3% of the 1103 sentences in the text. The remaining 19 complete failures constitute 1.7% of the sentences. Because 8 sentences in the German source text were found to have substantial errors resulting in poor or no translations, the quality assessment in Table 7.5 assumes a base of 1095 sentences.

In addition to the automatic scoring program to categorize the results as revealed in Tables 7.1-7.4, a staff linguist fluent in German, and with a technical background, scored the translations. In particular, the "successful" translations automatically assigned to Tables 7.1-7.2 and the "failures" assigned to Table 7.3 were investigated to determine their acceptability, if any. In general, the translations automatically categorized as "successful" were precisely that: 93% of the English target strings were well-formed, and were judged to convey the intended message. All of the 213 1-word direct translations assigned to Table 7.1 were judged acceptable; of the 780 successfully analyzed sentences assigned to Table 7.2, only 68, or 6.8% of all analyses, were judged erroneous in translation. There were various types of errors in the translations assigned to Table 7.3. In 13 cases the translations were perfectly good, except for the '|' characters METAL inserts to notify the post-editor that a phrasal translation has been performed; another 19 were very good phrasal translations, with at most one word requiring deletion or reposition; 59 were lesser-quality phrasal translations.

When investigating the causes of errors, it was determined that only a few basic problems were present. For example, half of the 68 erroneous translations where sentential analyses were produced resulted from failures in only four linguistic rules. In addition, 63% of the phrasal translations were necessitated by similar errors which prevented the achievement of a sentential analysis. It is usually the case in such a system as ours that a few errors in critical places will affect a disproportionate number of translations. Correspondingly, it is also true that correcting these few errors produces a large increase in reliability.

In the final analysis, there is no objective measure of translation quality. What seems most reasonable to us is an operational definition: the amount of time and effort required to revise the translations. Since post-editing is standard operating procedure for human translation, it seems reasonable to compare the post-editing costs for machine translation with those for human translation. Siemens AG informs us that post-editing human translations requires about 1/4 the time needed to perform the initial translation; further, that the human translation rate for material similar to that employed in this experiment is about four pages/day. From these data one can deduce that the expected post-editing time for a human translation of these 50 pages would be 25 hours. Using the various forms of output produced by our MT system, and an unsophisticated on-line editing program, an LRC staff linguist edited the reconstituted target text, reworking the phrasal translations and producing original translations where METAL failed to produce an acceptable translation. The time required was 15 hours. Together with the 2 hours required to prepare the original source text, editing time totalled 17 hours, less than what would be expected in the case of human translation. And while post-editing requires a highly-trained translator/editor, the preparation task, as discussed in Chapter VI, can be performed by a monolingual person with very little training. In addition, the use of one of the more sophisticated text editing programs should reduce preparation and revision time even further. Thus the total editing costs for material translated by METAL may be considerably lower than the equivalent for human editing. This constitutes a very significant advance in the state of the art in MT. No previous system has demonstrated editing requirements approaching those for human translation, to say nothing of reducing them.

### 7.3. Cost-Effectiveness Determination

Table 7.6 summarizes the machine costs entailed in performing the translations. Table 7.7 presents the total cost encountered during this experiment. Cost-effectiveness is of course determined by the cost of machine versus human translation, but for economic reasons one must include in the cost of machine translation an amortization of the costs of system development and maintenance, plus the cost of the necessary hardware. We assume an amortization period of 5 years, and a computer dedicated to MT and directly associated tasks (e.g., preparation and revision). We estimate the costs of producing, operating and maintaining the system as follows: a development period of 15 man-years at \$25,000 per man-year, \$400,000 for a suitable computer, 5 man-years per year for system operation and maintenance at \$25,000 per man-year, plus \$150,000 for a license fee. Expenses thus total \$1,550,000 over the 5-year period, or \$310,000 per year. Siemens AG, with vast experience

in translating highly technical material of this nature, reports a translation cost of around \$1 per line, including post-editing. At this rate, the machine must translate -- and humans must revise -- 310,000 lines per year to break even in 5 years. With 250 8-hour working days per year, the machine must translate 155 lines per hour, or about 18 words per minute. Two or perhaps three post-editors would be required for this volume of output. If this rate can be sustained, which seems reasonable given our results, MT costs would equal human translation costs during the amortization period and drop by a factor of 2.5 thereafter.

There are other economic factors that work to MT's advantage. The cost of human translation is going up by about 10% per annum, while machine costs are still dropping rapidly. More intensive application, e.g., working in multiple on-line shifts or translating in "batch" mode during off-hours, would ensure MT's profitability even during the amortization period. Moreover, translation into several languages from a single analysis would multiply MT's effectiveness, because most of the MT effort is expended in analyzing the input. Thus it would appear that our experiments have demonstrated both feasibility and cost-effectiveness for MT using METAL.

Ln	#S	Analysis Phase				Trans/Gen Phase				
		#I/S	#P/S	#R/S	PurePT	#Cells	FairPT	PureGT	#Cells	FairGT
1	213	0.0	0.0	0.0	0.01	2.0	0.01	0.25	128.5	0.34

Table 7.1  
 Direct Translations  
 Of 1-Word Items

Ln	#S	Analysis Phase					Trans/Gen Phase			
		#I/S	#P/S	#R/S	PurePT	#Cells	FairPT	PureGT	#Cells	FairGT
1	37	1.0	9.3	5.1	1.61	986.8	2.25	1.02	228.5	1.17
2	120	1.7	48.4	33.4	2.60	1951.7	3.88	1.04	521.0	1.38
3	108	1.5	37.5	22.6	2.80	1681.7	3.90	0.92	475.8	1.23
4	98	1.8	75.7	51.8	3.82	3016.9	5.79	1.52	772.0	2.02
5	77	7.1	140.2	86.8	6.00	5570.6	9.64	2.07	1034.0	2.74
6	47	3.5	214.4	138.3	8.89	8847.5	14.67	2.82	1069.9	3.52
7	44	4.0	261.5	170.5	10.39	10914.2	17.52	2.53	1130.4	3.27
8	51	6.4	242.7	143.0	16.05	13547.3	24.90	3.73	1506.6	4.72
9	37	6.6	215.2	123.4	17.16	11562.3	21.72	2.78	1531.9	3.78
10	32	3.7	317.6	205.8	17.69	14983.8	27.48	4.10	1701.5	5.21
11	25	7.5	426.1	223.1	31.18	28090.8	49.53	4.23	1788.2	5.40
12	34	8.6	416.1	239.1	25.40	21479.2	39.43	4.33	1963.3	5.61
13	31	8.8	424.3	242.1	28.60	27208.7	46.38	4.70	2080.7	6.06
14	4	8.0	625.7	362.2	41.04	36165.7	64.67	4.79	2189.5	6.22
15	9	18.4	1359.5	907.0	94.39	103042.0	161.71	7.03	2659.6	8.77
16	4	30.0	1428.2	866.5	77.79	82997.0	132.02	5.82	2896.0	7.71
17	4	23.0	1014.2	610.0	66.38	63314.0	107.75	5.83	2830.2	7.68
18	4	9.0	848.2	649.5	32.18	26147.2	49.27	7.53	2829.7	9.38
19	3	5.0	734.6	467.3	33.84	29329.3	53.00	3.82	2438.6	5.41
20	4	39.0	1099.5	667.5	60.29	58755.5	98.68	7.53	3154.2	9.59
21	4	8.0	1193.7	695.5	61.68	56980.2	98.91	7.98	2983.5	9.93
26	1	9.0	700.0	435.0	46.29	32342.0	67.42	11.35	4741.0	14.45
27	1	48.0	1012.0	558.0	67.72	55084.0	103.71	12.58	4767.0	15.69
34	1	60.0	2336.0	1764.0	251.20	326461.0	464.49	8.42	4927.0	11.64

Table 7.2  
Successful Analyses  
and Complete Translations

Ln - Length of item to be translated  
 #S - Number of sentences (items) of the given length  
 #I/S - Average number of interpretations/sentence  
 #P/S - Average number of phrases instantiated/sentence  
 #R/S - Average number of phrases rejected on semantic grounds  
 PurePT - Average pure parse time expended/sentence  
 #cells - Average number of cells (memory words) used/sentence  
 FairPT - Average fair parse time expended/sentence  
 PureGT - Average pure trans/gen time expended/sentence  
 FairGT - Average fair trans/gen time expended/sentence

Ln	#S	Analysis Phase					Trans/Gen Phase			
		#I/S	#P/S	#R/S	PurePT	#Cells	FairPT	PureGT	#Cells	FairGT
1	2	0.0	0.0	0.0	0.01	2.0	0.01	0.24	139.5	0.33
3	2	0.0	34.5	24.5	4.53	3242.0	6.65	1.55	494.5	1.87
4	1	0.0	114.0	68.0	4.68	4808.0	7.82	1.64	1076.0	2.34
5	1	0.0	65.0	42.0	2.80	2781.0	4.62	1.86	982.0	2.50
6	2	0.0	31.0	10.0	2.96	2154.5	4.37	0.81	860.0	1.37
7	5	0.0	119.0	77.3	7.11	5763.0	10.87	1.77	971.1	2.41
8	5	0.0	185.3	115.3	10.67	8571.0	16.27	3.59	1414.0	4.51
9	1	0.0	485.0	263.0	35.09	32636.0	56.42	2.62	1316.0	3.48
10	4	0.0	167.7	112.2	11.36	8364.5	16.83	3.67	1704.5	4.78
11	10	0.0	380.4	252.3	18.59	17747.6	30.19	4.58	1959.0	5.86
12	2	0.0	363.0	267.0	19.12	14914.5	28.87	9.82	2567.0	11.50
13	3	0.0	399.6	247.0	22.01	19924.6	35.02	5.78	2333.6	7.31
14	23	0.0	537.7	290.1	34.55	33526.0	56.46	5.65	2303.2	7.15
15	3	0.0	595.0	416.6	31.65	30566.0	51.62	8.66	3039.3	10.65
16	1	0.0	731.0	400.0	42.17	41119.0	69.04	5.24	2431.0	6.83
17	8	0.0	570.2	318.6	37.11	30873.1	57.28	7.36	2655.3	9.10
18	4	0.0	361.0	229.2	23.25	18549.2	35.37	6.45	3087.5	8.46
19	1	0.0	612.0	419.0	43.91	26375.0	61.14	6.17	2763.0	7.98
20	2	0.0	318.5	184.0	25.29	20120.0	38.43	6.69	3207.0	8.79
21	1	0.0	416.0	268.0	38.31	23722.0	53.80	8.99	3391.0	11.21
22	2	0.0	969.0	743.0	33.33	27124.5	51.05	9.87	3405.5	12.10
25	3	0.0	530.0	341.3	28.26	25878.6	45.17	6.70	3440.0	8.95
28	3	0.0	1871.3	1304.3	109.32	99886.0	174.58	10.99	4536.3	13.96
29	2	0.0	1176.5	834.5	60.05	54948.0	95.95	8.82	4563.5	11.80

Table 7.3  
Analysis Failures  
With Phrasal Translations

Ln - Length of item to be translated  
 #S - Number of sentences (items) of the given length  
 #I/S - Average number of interpretations/sentence  
 #P/S - Average number of phrases instantiated/sentence  
 #R/S - Average number of phrases rejected on semantic grounds  
 PurePT - Average pure parse time expended/sentence  
 #Cells - Average number of cells (memory words) used/sentence  
 FairPT - Average fair parse time expended/sentence  
 PureGT - Average pure trans/gen time expended/sentence  
 FairGT - Average fair trans/gen time expended/sentence

Ln	#S	Analysis Phase					Trans/Gen Phase			
		#I/S	#P/S	#R/S	PurePT	#Cells	FairPT	PureGT	#Cells	FairGT
13	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
21	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
29	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
30	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
31	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
32	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
34	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
36	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
37	1	0.0	887.0	583.0	42.29	37838.0	67.01	0.05	2.0	0.05
43	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 7.4  
 Complete Failures  
 No Analysis or Generation

Ln - Length of item to be translated

#S - Number of sentences (items) of the given length

#I/S - Average number of interpretations/sentence

#P/S - Average number of phrases instantiated/sentence

#R/S - Average number of phrases rejected on semantic grounds

PurePT - Average pure parse time expended/sentence

#Cells - Average number of cells (memory words) used/sentence

FairPT - Average fair parse time expended/sentence

PureGT - Average pure trans/gen time expended/sentence

FairGT - Average fair trans/gen time expended/sentence

Original document:	1103 sentences	
	50 pages	
	22.1 sent/page	
Major source errors:	8 sentences	
Computation base:	1095 sentences	
Analyzed:	985 sentences	90.0%
Excellent trans.	917 sentences	83.7%
Inadequate trans.	68 sentences	6.2%
Not analyzed:	110 sentences	10.0%
Excellent phrasal	32 sentences	2.9%
Lesser phrasal	59 sentences	5.4%
No trans.	19 sentences	1.7%
Total errors:	178 sentences	16.3%
Good phrasal	32 sentences	(18%)
Lesser phrasal	59 sentences	(33%)
Inadequate trans.	68 sentences	(38%)
No trans.	19 sentences	(11%)
Sentences:	22.06/page	
Good trans. @83.7%	18.46/page	
Errorful trans. @16.3%	3.60/page	

Table 7.5  
Quality Assessment

	Table 7.1	Table 7.2	Table 7.3
No. sentences	213	780	91
No. words	22133	4818	1258
Per sent.	1.0	6.2	13.8
Total parse time	2.13 sec.	4.21 hrs.	1.12 hrs.
Per sent.	0.01 sec.	19.44 sec.	44.24 sec.
Per word	0.01 sec.	3.15 sec.	3.20 sec.
Total trans/gen time	72.42 sec.	0.68 hrs.	0.18 hrs.
Per sent.	0.34 sec.	3.15 sec.	6.99 sec.
Per word	0.34 sec.	0.51 sec.	0.51 sec.
Total trans. time	74.55 sec.	4.89 hrs.	1.30 hrs.
Per sent.	0.35 sec.	22.59 sec.	51.23 sec.
Per word	0.35 sec.	3.66 sec.	3.71 sec.
Total trans. cost	\$3.07	\$729.47	\$192.95
Per sent.	\$0.014	\$ 0.93	\$ 2.07
Per word	\$0.014	\$ 0.151	\$ 0.153

Table 7.6  
 Summary of Times and Costs  
 Where Translations Were Provided

	Machine Cost	+	Human Time
Pre-editing	\$ 9	+	2 hours
Dissociation	\$ 12		
Translation	\$1090		
Reconstitution	\$ 9		
Post-editing	\$ 15	+	15 hours
	-----		-----
Total	\$1135	+	17 hours

Table 7.7  
 Total Machine + Human Costs  
 Encountered During Experiment

Pre-editing: Marking sentence boundaries and resolving occasional format difficulties in the source text.

Dissociation: Splitting the computer-typeset source text into sentences to be translated and other material (e.g., flowchart boxes) to be maintained as is.

Translation: Analyzing source- and synthesizing target-sentences.

Reconstitution: Constructing a target text in typeset form, using target- in place of the original source-sentences.

Post-editing: Validation and, as necessary, emendation of the target text, plus any required (re)translation of the source.

## 8. Conclusions

### 8.1. Implications of Achieving Fully Automatic High Quality Translation

Many claims have been made for MT systems, and one may well ask whether those for METAL have been overstated. A convincing answer may be obtained by examining its design and performance.

The design is in accordance with strict linguistic theory. Linguists have long assumed the necessity of three components for a grammar: one to deal with the segments, the phonological component; one to deal with meaningful elements and their arrangements, the syntactic component; one to deal with the relations of these elements to the outside world, the semantic component. Specialists in communication theory, following Charles Sanders Peirce, propose three broader components: the syntactic, semantic and pragmatic. Only the last of these has not been identified by earlier linguistic theory; it deals with the relations between language and users. These facets, which must be included in a fully developed system of linguistic description (Bar-Hillel, RADC-TR-71-295 I.93-94), have all been accounted for in the design of METAL.

Earlier sections of this report have discussed the treatment of the syntactic and semantic components. A brief comment on the pragmatic component may be added here. Technical writing imposes a different relationship between language and users than do other uses of language. Lexical items are more carefully restricted in application. Syntactic patterns are strictly regulated. Pragmatic facts can accordingly be determined with regard to technical writing. Because technical writing is relatively straightforward, technical materials are far more amenable to pragmatic treatment than are many other types of language. For this reason too, descriptions of technical writing are simpler than are descriptions of less restricted uses of language. Additional features will be required to achieve adequate analysis of general materials.

Advances in software and hardware have made the treatment of numerous features manageable, as the sections above on the program component indicate. Further improvements in design will bring greater speed of translation even as additional features are introduced.

### 8.2. Operational Requirements

The aims and funding of the research yielding the METAL system permitted only an experimental version. Past skepticism about the viability of fully automatic high quality translation required demonstration of its feasibility on the one hand and of economic advantages on the other. Since both have been demonstrated, METAL can now be expanded to an operational system.

For this expansion the lexicon of any given field must be described as thoroughly as that of German telecommunications has been. There is in principle no limit on type of vocabulary which can be handled. Nor are there limits on syntax. New syntactic patterns will inevitably occur. They will require modification of the current rules, a process calling for highly

capable linguists but in no way impossible. It might well be noted that METAL was developed under very trying conditions for both software and hardware.

### 8.3. Contributions of Further Research

Linguistic analysis in the past has been carried out as was research in the physical sciences two centuries ago. Scholars and their students analyzed whatever data interested them. The process led to important discoveries, much as did the activity of chemists and physicists at their work-benches. But the linguistic data that could be analyzed in the past were infinitesimal compared with data currently available. Results of research based on such data are eloquently discussed by one of the leading French linguists (Maurice Gross, *On the Failure of Generative Grammar*, *LANGUAGE* 55 [1979] 859-885). Operational MT will bring the possibilities for improved linguistic research discussed by him into realization.

Benefits for our understanding of language might be stated at length. For brevity, illustrations will be given for only a few points of Bar-Hillel's summary (RADC-TR-71-295 I.93-94). Ongoing analysis of large quantities of data will provide "information about speech act conditions and conversation rules," resulting in the production of a "theory of discourse." It will also provide means for testing theories of natural logic by which one may make "judgements on the success of an argument" or by which "the appropriateness of elements in conversation can be deduced." Linguists have been hampered in dealing with these topics, and may well be troubled by the analysis of huge quantities of data. Yet as Gross points out, size of these data "would be considerably smaller than the number of pictures taken daily from bubble chambers and analyzed by physicists" (1979:879). In short, linguistics will enter a new phase of research.

The benefits resulting from finer analysis of texts also need little elaboration. MT is only the first step in data processing. If language can be analyzed for translation, it can also be treated for content. Use of semantic features, like those incorporated in METAL, will make possible great advances in access to data. The simple methods now in use for data retrieval and indexing compare in sophistication with first-generation MT. More advanced applications, such as fact retrieval, will follow. As such applications are developed linguists concerned with translation will receive far greater recognition than they have in the past, much as chemists, physicists, and other specialists were accorded recognition when they moved from individual tasks to more theoretical ones on the one hand and more managerial tasks on the other. Translators will finally be recognized for their professional competence rather than merely as slightly more skilled than any bilingual speakers.

Finally, a society with more than half its members professionally involved in communication needs appropriate means to process documents rapidly and efficiently. METAL illustrates the capability of linguistics and computer sciences to provide such means.

THE METAL SYSTEM  
Vol. 2: Appendices

### Table of Contents

APPENDIX A-1: Lexical Features and Values - Introduction	1-5
adjectives.....	6-21
adverbials.....	22-51
connective	22-25
degree	26-29
locative	30-33
manner	34-37
particles	38-41
prefixes/adpreps	42-44
preverbs	45-47
temporals	48-51
conjunctions.....	52-53
determiners &.....	54-59
predeterminers.....	60-63
nouns.....	64-76
prepositions.....	77-81
pronouns.....	82-87
verbs &.....	88-103
modals/auxiliaries.....	104-111
APPENDIX A-2: Semantic Type.....	112-121
APPENDIX A-3: Argument Roles.....	122-132
APPENDIX A-4: Provenience Tags.....	133-147
APPENDIX B: Rules.....	148-155
APPENDIX C: Transformations.....	156-157
APPENDIX D: Case Frames.....	158-162

## APPENDIX A-1

### LRC Lexical Data Bases for MT

We record ('code') lexical information into matrices which will become 'entries' used in mechanical translation, currently from German to English. The skeletal matrices for potential lexical rules consist of lists of general attributes ('features') which characterize the various lexical categories. These lexical categories roughly correspond to traditional parts of speech, as can be seen from the list under "CAT", below.

For each feature, particular 'values' pertinent to the word-sense are specified by a human coder or assigned by the computer. The value for onset (ON), the consonantal or vocalic quality of the sound a word begins with, for instance, can be automatically determined for most words, but the grammatical case governed by a given verb requires human knowledge.

The features used by the system in parsing and generating natural language can be considered in two groups, systemic features and specific features.

Systemic features, identified by their three-letter acronyms, are a required component of every lexical entry regardless of language or particular part of speech. Since they are universal to the system, these features are described here in the introduction rather than repeated in the write-up for each lexical category. (For reasons of economy the descriptions of very lengthy value sets which apply to features in a number of lexical categories are listed separately in Appendix A-2, A-3, and A-4.) Two of the universal system features, CAN <canonical form> and ALO <allomorph>, are also discussed in certain of the lexical category descriptions since the choice of values is sometimes conditioned by the paradigmatic patterning of the lexical entry.

Language- and lexical-category-specific features, identified by having two-letter acronyms, are determined by the language and the part-of-speech (CAT) to which the lexical entry belongs. Thus they cannot be considered universal. (In fact, categories such as the PRFX <verb 'prefix'> in German and English do not require any language- or category-specific features at all and thus only carry the obligatory systemic features.) The specific, as opposed to systemic, features carry morphological, syntactic, and semantic information necessary to ensure quality translation. Although the same feature may be used in more than one language, or more than one lexical category within a single language, neither the use of the feature within the system nor the relevant values are necessarily the same. The reader is thus referred to the individual CAT <lexical category> sections for a description of these features.

The nine features common to all lexical categories are:

CAN	canonical form
ALO	allomorph
CAT	lexical category
PLC	placement
PRF	preference
LEX	lexical collocation
SNS	sense number
CNO	concept number
TAG	provenience tag

CAT <lexical category> refers to the class of lexical entries in which the word or affix participates. This feature corresponds loosely to the traditional concept, part-of-speech. The values are:

AST	= adjective stem
NST	= noun stem
VST	= verb stem
MDX*	= modal/auxiliary
DST	= determiner stem (German only)
DET	= determiner
PREDET	= predeterminer
LOC	= locative adverb
MAN	= manner adverb
PREVB	= (adverbial) preverb
TMP	= temporal adverb
DEG	= degree adverb
PRT	= adverbial particle
CON	= connective adverb
PRFX	= adverbial 'prefix'
CONJ	= conjunction
PREP	= preposition
PRN	= pronoun
N-FLEX**	= noun inflection
A-FLEX**	= adjective inflection
D-FLEX**	= determiner inflection
V-FLEX**	= verb inflection
N-INFIX**	= noun infix
V-INFIX**	= verb infix
PNCT**	= punctuation

\*Certain auxiliaries such as "sein", "haben", "werden", "be", "have", "will", form their own categories, but for convenience in discussion we refer to them as 'MDX'

\*\*As the inflectional and punctuation categories utilize only the standard set of systemic features, they are not included among the following write-ups.

CAN <canonical form> corresponds roughly to the concept of 'morpheme' or 'lexeme'. It is a single form which is used to represent all the variations a word may take due to paradigmatic or stylistic manipulation. Usually the form most likely to be included in a dictionary is selected, such as nominative singular for nouns, the infinitive for verbs, etc. Multi-word entries are handled variously, according to category. Particular instructions for selection of the CAN may be found in the write-ups for certain lexical categories.

When two different morphemes have identical canonical forms, a number is added to the alphabetic sequence to make them distinct, e.g., FAST1, FAST2, etc. Divergent senses of the same word may be further distinguished by differing values for SNS <sense number>, as described below.

During translation, transfer keys on the CAN for lexical items. In actual entry format the CAN is not marked by the acronym but by its position as the initial element, which makes the letters "CAN" superfluous.

AL0 <allomorph> represents the actual surface string. If it is a phrase, for instance, blanks are used where expected. The concept corresponds to that traditionally associated with 'allomorph'. Thus variants of a word such as "mice"/"mouse", "man"/"men", or "be"/"am"/"was"/"were" would be assigned to AL0 according to the actual string covered by the entry but all of the AL0's for a paradigm must have the same CAN <canonical form>. Additional details and examples may be found in the individual write-ups.

PLC <placement> indicates the position of a morph with respect to other morphs contained within the same word. The values are:

WF = word final (the entry cannot be followed by another morph: it must be followed by a word boundary such as blank or punctuation)

WI = word initial (the entry cannot be preceded by another morph: it must be preceded by a blank, punctuation, or sentence initial position)

NI = non-initial (the entry must not be the initial morph in the word: it cannot be preceded by a word boundary but must be preceded by another morph; suffixes are typical examples)

NF = non-final (the entry must not be the final morph in the word: it cannot be followed by a word boundary but must be followed by another morph; prefixes are typical examples)

Note that these values frequently occur in clusters. Inflectional endings must be suffixed to another morph; in addition, they must be the final element in the word (WF NI). A derivational prefix, on the other hand, must be word initial and be affixed to the front of another morph (WI NF). An infix might require both preceding and following morphs (NI NF), whereas a pronoun may only occur as an unbound form (WI WF).

PRF <preference> is used to prevent multiple translations where they are not justified. It does so by giving certain readings preferred treatment over others. Thus a noun compound which is entered in the lexicon as such is given preference over a reading which finds its individual sub-components and analyzes them. The system does so by means of a weighting factor indicating how highly valued a particular analysis may be. Values are numeric, with 1 indicating no preference; 2 multiplying the weighting factor by 2 ("twice as good"); and 0.5 dividing the value by 2 ("only half as good"). The weights of the individual nodes are carried on up through the sentence. At the time of transfer, the tree with the highest value is attempted first; should it fail, the next highest which succeeds is selected.

LEX <lexical collocation> Morphemes which occur in idiomatic or quasi-idiomatic expressions whose elements may be discontinuous are marked by this feature. It 'triggers' the system to look for the related elements so that they may be translated as a unit rather than as a sum of the individual literal meanings.

When the elements of such a locution are invariably contiguous, it may be entered in the lexicon as a single entry containing blanks, with appropriate preference rating (PRF). This is more economical since it avoids the necessity of applying the special idiom look-up and reordering routines.

(Note that even in phrases whose words would ordinarily be contiguous, inflectional endings such as past tense may intervene, necessitating entry as a lexical collocation because of the internal variable.) The values for LEX are:

T = true - the element is used in an expression which includes internal variables (i.e., is discontinuous)

NIL = lexical collocation look-up is not applicable

SNS <sense number> is used to indicate different senses or meanings of the same string. The values are numbers, often preceded by a letter specifying a particular reference dictionary, and are assigned by the system or by the coder as a means for distinguishing homographic entries.

CNO <concept number> groups semantically related words together, both within a single language and across several languages. Thus the verb "compute", the nouns "computer" and "computation", and the adjective "computable" would share the same concept number, the particular translation being the part of speech required by the sentential context.

TAG <area of provenience tag> indicates the discipline(s) in which a particular word or word sense is most likely to be used. Often a term has both a technical and a general meaning. The translation would differ depending upon the discipline being translated, which is keyed by TAG. (See Appendix A-4 for lists of TAG values.)

## GERMAN ADJECTIVE FEATURES

CAT = AST

*CAN	canonical form
*ALO	allomorph
CL	inflectional class
{CA}	grammatical case
{NU}	grammatical number
{GD}	grammatical gender
{IN}	inflection
{DG}	degree
{PO}	position
CP	capitalization
RM	'case' role of modificand
TM	semantic type of modificand
MM	surface marker of modificand
FM	syntactic form of modificand
RC	'case' role of complement
TC	semantic type of complement
MC	surface marker of complement
FC	syntactic form of complement
SP	occurs in special frame
AF	grammatical form of adjective
HG	homographic with
DF	derivable from
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix. Those set off by {} are brought in by means of the CL <inflectional class> feature during processing rather than coded separately.)

DISCUSSION: When coding a German adjective it is first necessary to determine how many stem forms the adjective has. This is because multiple forms will require separate lexical entries. An adjective such as "schlecht", for example, would require only one entry, since the

positive, comparative, and superlative forms ("schlecht", "schlechter", "schlechtest") all make use of the same stem, "schlecht". On the other hand, an adjective such as "alt", which may be umlauted ("alt", "aelter", "aelttest"), has two stem forms and would require two separate lexical entries, one for "alt" and one for "aelt".

Although not normally the case, a German adjective may have as many as four distinct stem forms:

Example	Stem	Degree	Syntactic Position
der Berg ist hoch	hoch	positive	predicative only
der hohe Berg	hoh	positive	attributive only
der Berg ist hoeh	hoeh	comparative	predicative or
der hoehere Berg			attributive
der Berg ist am hoechsten	hoech	superlative	predicative or
der hoechste Berg			attributive

Each distinct stem form will be the basis for a separate lexical entry. Each of the four entries above would have "hoch" as its CAN <canonical form> and the particular stem as its ALO <allomorph> (as illustrated under ALO below).

CAN <canonical form> is usually the positive predicative form of the adjective. If the adjective cannot be used predicatively, e.g., "hiesig" or "ober", its attributive stem is used without endings. If alternative spellings of the predicative form are possible, e.g., "feig/feige", the shorter form is chosen.

ALO <allomorph> does not need to be coded unless a stem form differs from the canonical form of the word. The adjective "hoch", as mentioned above, would require four separate entries with the following values for CAN <canonical form> and ALO:

CAN	hoch	hoch	hoch	hoch
ALO	hoh	hoch	hoeh	hoech

Adjectives such as "alt" and "dunkel" would require two entries each:

CAN	alt	alt
ALO	alt	aelt
CAN	dunkel	dunkel
ALO	dunkel	dunkl

CL <inflectional class> specifies inflectional affixes which may occur with a particular adjective stem. The values are a combination of a degree-class identifier plus a marker designator, which serve to identify the endings for the ALO <allomorph> in each of the following categories:

- (1) positive attributive form (values begin with PA-)
- (2) positive predicative form (values begin with PP-)
- (3) comparative form (value begins with C-)
- (4) superlative form (values begins with S-)

In coding, each ALO <allomorph> should be assigned at least one and not more than five values for CL <inflectional class>, depending upon how many categories are relevant to the particular stem form being coded. For example, the adjective "noetig" would receive the following values for CL:

ALO	CL
noetig	(PA-N PP-Ø C-ER S-ST)

Positive Attributive Values indicate whether the ALO may be used attributively and, if so, what endings it takes.

PA-N = used attributively with normal adjective endings  
(ein rotes Auto, die schoene Frau)

PA-Ø = used attributively without endings  
(ein rosa Kleid, das Muenchner Bier)

(no value) = not used attributively (hoch, groess)

Positive Predicative Values indicate whether the ALO <allomorph> may be used predicatively and, if so, whether it takes a zero or -e ending.

PP-Ø PP-E = used predicatively with zero or -e ending  
(der Mann ist feig/feige)

PP-Ø = used predicatively with zero ending only  
(das Kleid ist schoen)

(no value) = not used predicatively (hiesig, ober, link)

Comparative Values indicate whether the ALO <allomorph> has a comparative form.

C-ER = used comparatively with -er/-er- affix  
(die schoenere Frau, dieser Mann ist aelter)  
C-Ø = used comparatively with no affix  
(eine bessere Loesung, dieser Berg ist hoher)  
(no value) = not used comparatively (ober, gross, gut)

Superlative Values indicate whether the ALO <allomorph> has a superlative form and, if so, which affix is used to build it.

S-ST S-EST = used superlatively with -st- or -est- affix  
(die stupidste/stupideste Arbeit)  
S-EST = used superlatively with -est- affix only  
(das kuerzeste Stueck, der Diamant ist am haertesten)  
S-ST = used superlatively with -st- affix only  
(das schaerfste Messer, der Mann ist am feigsten)  
S-T = used superlatively with -t- affix only  
(die groesste Freude, dieses Buch ist am groessten)  
S-Ø = used superlatively with no affix  
(die meisten Studenten, die Loesung ist am besten)  
(no value) = not used superlatively (rosa, Muenchner, gross, hoch)

CA <grammatical case> has the traditional values:

N = nominative  
G = genitive  
D = dative  
A = accusative

NU <grammatical number>

SG = singular  
PL = plural

GD <grammatical gender>

M = masculine  
F = feminine  
N = neuter

IN <grammatical inflection> is coded to allow for various combinations of prenominal modifiers.

ST = strong  
WK = weak

DG <degree of comparison>

POS = positive  
COM = comparative  
SUP = superlative

PO <syntactic position> is required since some adjectives take different inflectional affixes depending upon whether they occur before a noun (attributive) or after the copula (predicative).

ATR = attributive  
PRD = predicative

CP <capitalization> is used to indicate adjectives that are derivatives of proper nouns.

UC = upper case (proper noun derivative)  
LC = lower case (not a proper noun derivative)

RM <role of modificand> refers to the semantic relation of the modificand. For the majority of adjectives entered in the lexicon as adjective stems, this feature is superfluous as it is not used by subsequent syntactic rules (most adjectives can be used with almost any noun). However, deverbal nouns (which are created from verb stems by the grammar) do place restrictions upon the function of their modificands, present participles generally modifying their agents and past participles more frequently modifying their targets. The values are those for central case roles (see Appendix A-3 for a complete list), in particular:

ACT = agent  
TAR = target  
INS = instrument

TM <semantic type of modificand> specifies the semantic character of nouns which can be modified by the adjective. For example, the adjective "zornig" can refer to something which is animate (die zornige Frau) or to an abstraction (der zornige Blick) but not to an inanimate object (\* der zornige Stein). Values for delineating such characteristics are to be found in the semantic type matrix in Appendix A-2.

MM <modificand marker> This feature is redundant in that adjective forms do not restrict the case or prepositional markers of the nouns they modify. It is thus not coded for most adjectives. However, certain adjectives which are derived by the grammars from verbs may carry such restrictions from their stem forms. This information is retained for ease in transformation should an adjective form not be selected for translation. See MA <mark of argument> in the Verb feature description for additional information.

FM <syntactic form of modificand> specifies the constituent classes which may be modified by the adjective.

NP = adjective may modify a noun phrase [DEFAULT]  
CP = adjective may modify a complement phrase or clause

RC <role of complement> refers to the 'case grammar' roles which the complement of the adjective can play. Values for this feature are to be found in a separate list (see Appendix A-3).

TC <semantic type of complement> indicates semantic restrictions on the argument which functions as complement to the adjective. The values are those of the semantic type matrix (see Appendix A-2).

MC <complement marker> may take the form of a grammatical case, a preposition, or a grammatical construction.

G = genitive

D = dative  
A = accusative  
— = (any prepositions, the value being their canonical form)  
(Sie ist auf eine Unterstuetzung angewiesen.)  
TH = "dass" clause  
(Es ist wichtig, dass der Wagen repariert wird.)  
FT = infinitive complement  
(Es ist notwendig, den Motor auszutauschen.)  
WH = subordinate question  
(Es ist aber fraglich, ob wir einen finden werden.)

FC <syntactic form of complement> specifies the constituent which functions as the complement or object of the adjective.

PP = prepositional phrase  
CP = complement phrase or clause

SP <special grammatical frames> are special constructions into which certain adjectives may be inserted.

PD = personal dative frame: "Mir ist \_\_\_\_."  
(Mir ist kalt/warm/schlecht/angst und bange)  
ID = impersonal dative frame: "Das ist mir \_\_\_\_."  
(Das ist mir bekannt/neu/gleich/egal)

AF <grammatical form of adjective> is used, for German, to indicate a deverbal adjective.

PAPL = past participle (der gedeckte Tisch)  
PRPL = present participle (fliessendes Wasser)

HG <homograph> Adjectives which have the same spelling as another part of speech are given values indicating the syntactic class or the form matched.

NO = noun  
VB = verb  
OTR = other part of speech

Although the above feature is not used by the system at present, it may be included in the current lexical entries as the basis for disambiguation rules to be incorporated into the finished system.

DF <derivable from> will be assigned to derived adjectives, usually deverbal forms such as participles, but also certain adjectives derived from noun stems or other adjectives. This feature is to particularize the underlying word in the event that some of its grammatical characteristics may have been carried along during derivation.

VI = intransitive verb  
VT = transitive verb  
VR = reflexive verb  
ADJ = another adjective  
NO = noun

## ENGLISH ADJECTIVE FEATURES

CAT = AST

*CAN	canonical form
*AL0	allomorph
CL	inflectional class
{DG}	degree
NU	grammatical number
CP	capitalization
RM	'case' role of modificand
TM	semantic type of modificand
MM	surface marker of modificand
FM	syntactic form of modificand
RC	'case' role of complement
TC	semantic type of complement
MC	surface marker of complement
FC	syntactic form of complement
PO	position
SQ	sequence among adjectives
NN	number when nominalized
LY	to derive related "-ly" adverb
AF	grammatical form of adjective
HG	homographic with
DF	derivable from
ON	onset quality
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix. Those set off by {} are brought in by means of the CL <inflectional class> feature during processing rather than coded separately.)

CAN is a quasi-canonical form of the word used to identify related lexical entries. For adjectives it is typically the positive form.

ALO <allomorph> is, ordinarily, the sequence of letters that remains the same whether the word is positive, comparative, superlative, or takes the "-ly" as an adverb. E.g., the upper case portions of the following words--

CAPABLE	SNUG	CLOSE	READY	FULL
more CAPABLE	SNUGger	CLOSER	READier	FULLer
most CAPABLE	SNUGgest	CLOSEst	READiest	FULLest
CAPABLY	SNUGly	CLOSEly	READily	FULLy

(The exceptions are largely the irregular adjectives such as "good/better/best", for which each form is coded in a separate entry with "good" as the CAN <canonical form> and "good", "better", and "best" as the ALO's.) The examples shown under CL <inflectional class> illustrate varieties of ALO's.

CL <inflectional class> denotes the set of inflectional endings (including the zero affix) which apply to the ALO <allomorph>. The values specify the affixes for the positive, comparative, and superlative forms of the word. A zero signifies that no affix is required; absence of a P-, C-, or S-value signifies that the ALO <allomorph> does not occur in that degree.

Positive	Comparative	Superlative
P-Ø	C-Ø	S-Ø
P-E	C-ER	S-EST
P-Y	C-BER	S-BEST
	C-DER	S-DEST
	C-GER	S-GEST
	C-IER	S-IEST
	C-MER	S-MEST
	C-NER	S-NEST
	C-TER	S-TEST
	C-THER	S-THEST
	C-MORE	S-MOST
	C-E+MORE	S-E+MOST

Although nearly 400 combinations are theoretically possible, less than a score have actually been found to occur. These are shown on the following page. [The hyphen, used to indicate that the example word is truncated, is not coded.]

Examples	Value clusters
low, vast, few, full	CL(P-Ø) C-ER S-EST)
simpl-, abl-, clos-, whit-	CL(P-E C-ER S-EST)
glib, drab	CL(P-Ø) C-BER S-BEST)
mad, bald	CL(P-Ø) C-DER S-DEST)
snug, big	CL(P-Ø) C-GER S-GEST)
slim, dim	CL(P-Ø) C-MER S-MEST)
thin, tan	CL(P-Ø) C-NER S-NEST)
hot, wet	CL(P-Ø) C-TER S-TEST)
far	CL(P-Ø) C-THER S-THEST)
read-, lowl-	CL(P-Y C-IER S-iest)
capabl-, irascibl-	CL(P-E C-E+MORE S-E+MOST)
photographic, Greek, lateral, ablaze	CL(P-Ø) C-MORE S-MOST)
less (1)	CL(P-Ø) C-ER)
whol-, singl-	CL(P-E)
more, better, worse, less (2)	CL(C-Ø)
first, most, least, utmost, last, best, worst	CL(S-Ø)
up, above, adrift, absent, ultimate, bad, latter, adoptive, ablative, well-advised, accelerando, damn, abdominal, good, much, many	CL(P-Ø)

DG <degree> is in brackets above because it is introduced by the CL <inflectional class> value selected rather than coded in the lexicon.

POS = positive  
COM = comparative  
SUP = superlative

NU <grammatical number> of the modificand is specified when that is restricted to just singular or just plural.

SG = singular only ("solitary" [i.e., "sole"])  
PL = plural only ("various", "numerous")

CP <capitalization> marks those adjectives which are likely to be capitalized in situations other than at the beginning of a sentence or in a title.

UC = upper case ("French", "Texan")  
UC LC = upper case / lower case ("Roman"/"roman")  
LC = lower case [DEFAULT]

RM <'case' role of the modificand> The values are the same as for the other 'role' features. (See Appendix A-3)

TM <semantic type of modificand> takes values from the noun feature TY <semantic type>. As at times the same semantic values do not apply to an adjective in both the pre-modificand and post-copula positions, the satisfaction of this feature may necessitate multiple entries.

MM <marker of modificand>

TH = fits both frames, "it was ADJ that this happened",  
and "that this happened was ADJ"  
(e.g., "it was unfortunate that this happened",  
"that this happened was unfortunate")

FT = fits both frames, "it was ADJ for NOMINAL to VERB",  
and "for NOMINAL to VERB was ADJ"  
(e.g., "it was unusual for Regina to faint",  
"for Regina to faint was unusual")

PI = fits both frames, "NOMINAL's VERBing was ADJ" and  
"the fact that NOMINAL VERBed was ADJ"  
(e.g., "Bob's playing was remarkable"  
"the fact that Bob played was remarkable")

U = unmarked (for NP modificand)

FM <syntactic form of modificand>

NP = noun or noun phrase  
CP = complement phrase or clause

RC <role of complement> refers to Fillmorean 'case' roles that indicate the semantic function(s) of the type(s) of prepositional phrase(s) closely associated with the adjective. Such phrases are especially common with deverbal adjectives, but they also occur with non-derived adjectives, e.g.,

necessary for something  
deficient in something

The value of RC is the functional role of the following prepositional phrase. For a list of the possible values, see Appendix A-3.

TC <semantic type of complement> See semantic type matrix, Appendix A-2.

MC <syntactic marker of complement> takes the form of a preposition which typically follows the adjective (e.g., "absent + from", "happy + about"). Such prepositions mark predicate complement constructions similar to those found with verbal expressions, e.g., "Jane talked about the party", "Jane was vociferous about the party". (If the preposition invariably occurs or there are instances in which the object of the preposition can only be a gerund, accommodation for this will need to be made in the coding.) A prepositional value for this feature is always to be associated with the PO <position> values AFTCOP and AFTSUB.

\_ = (appropriate preposition)

TH = "that"-complement  
Fits the frame "she was ADJ that something happened"  
(e.g., "she was glad that they succeeded")

FT = "for...to" complement  
Fits the frame "she was ADJ for somebody to do it"  
or "she was ADJ to do it"  
(e.g., "she was willing for them to leave")  
"she was ready to go home"

WH = "wh"-word complement  
Fits the frame "she was ADJ wh\_ it was"  
(e.g., "she was uncertain who/where/why/etc. it was")

FC <form of complement> refers to the syntactic constituent which may function as a complement.

PP = prepositional phrase  
CP = complement phrase or clause

PO <position> of the unmodified adjective in relation to the other constituents in the clause. No value is coded if only the usual BEFMOD and AFTCOP apply.

BEFMOD = before its modificand

AFTCOP = after copula

AFTSUB = after the subject

MSR = can occur postposed in a construction of the type "five miles long", "ten man strong". [Note that the noun in such constructions cannot function as the subject of a sentence. Higher level rules will be needed to account for other circumstances in which a modified adjective may-- or even must-- follow its modicand (e.g., "paint the town red", "a man angry enough to...", "a terminal-specific program".]

SQ <sequence> notes the slot the word would ordinarily occupy in a string of prenominal modifiers, as in--

"all (the) first fifty huge beautiful  
tapering vermilion Minoan pillars"  
or  
"the next six short hairless elderly  
green Martian astronauts"

Most adjectives fall into a quality/characteristic category (e.g., "beautiful", "hairless") which is the default value. So far, ten kinds of modifiers have been noted which take a typical position preceding a noun modicand. [Certain extremely common adjectives like "little" and "old" do not seem to sequence consistently. Ordinarily, however, in both German and English, a forward shift in sequence confers emphasis (e.g., "Persian illuminated manuscripts", which focuses on the provenience as "illuminated Persian manuscripts" does not).]

PDT = predeterminers, q.v.

DET = determiners (which include indefinite adjectives such as "some" and "many"), q.v.

SEQ = 'sequencers' such as ordinal numerals or "next"

CNT = 'count' words (i.e., cardinal numerals)

SIZ = size

SHP = shape

AGE = age

CLR = color

OTR = other characteristic [DEFAULT]

PTI = personal title ("Mrs.", "Lord", "Comrade")

NN <number when nominalized> refers to grammatical number where only a singular or only a plural verb can be used with the word when it is nominalized by "the". (N.B., the nominal must not take a plural suffix.) Since in some context or another almost any adjective can be so nominalized, we will rely upon reference sources to determine which adjectives to code this feature for.

SG = typically only singular ("the departed", "the Almighty", "the sublime")

PL = typically only plural ("the dead", "the mighty")

LY <to derive related adverb> specifies the affix by which a semantically related adverb, if such exists, may be formed from the AL0 <allomorph> of the adjective. [The AL0's given as examples here are from the group used to illustrate the inflectional classes. Many of those, of course, do not form derivational adverbs (e.g., "Greek", "accelerando", "few"), while others ("low", "adrift", "farther", "less", "best") are homographic with their related adverb.]

Y = ("simpl-", "abl-", "singl-", "full", "capabl-", "irascibl-")

LY = (\*\*"first", \*\*"most", "whol-", "vast", "bald" [=unadorned], "dim", "thin", "glib", "drab", "wet", "hot" [=angry], "mad", "snug", "ultimate", "adoptive", "bad", "lateral", "abdominal")

ELY = ("clos-", "whit-")

ILY = ("read-")

ALLY = ("photographic")

Ø = ("low", "adrift", "less", \*\*"first", \*\*"most")

\*\* "Firstly" and "mostly" exist, as well as "first" and "most".

Although not presently being used by the system, derivational rules could permit multiple forms to be derived from a single lexical entry, saving storage space in the lexicon. Also, inclusion of such correspondences may allow flexibility in translation when the part of speech differs from one language to the other.

AF <grammatical form of the adjective>

N+ED = noun with "-ed" meaning "having" ("bearded", "red-eyed", "experienced", "figured")

PRPL = ends in "-ing" and could be confused with a true verb in participle form ("convincing", "becoming", "acting", "cutting", "engaging")

PAPL = related to a past participle ("distinguished", "accomplished", "collected", "decided")

HG <homograph> marks adjectives which have the same spelling as another part of speech. The values indicate the syntactic class of the form matched.

NO = noun other than gerund ("content", "dark", "African")

INF = infinitive ("direct", "equal", "content", "desert")

ADV = adverb ("far", "more", "worse", "deep", "farther", "farthest", "last", "up", "above", "first", "best", "adrift", "less", "least", "low", "better", "most")

OTR = other ("left")

Participial adjectives which occur in the reference dictionary will be coded on the assumption that they possess some semantic (e.g., "accomplished") or syntactic characteristic which distinguishes them as true adjectives (i.e., can be modified by "very", as in "very engaging").

DF <derivable from> is used for information concerning derivationally-underlying nouns or verbs in case the adjective may have 'inherited' some of their grammatical characteristics. It also records the existence of any synonymous, shorter form of the adjective (e.g., "algebraic/algebraical").

ADJ = another adjective

NO = noun

VI = intransitive verb

VR = reflexive verb

VT = transitive verb

ON <onset> or quality of initial sound, is required for selection of "a" or "an". Except for words beginning with h- or u-, the value is highly predictable from the spelling and so can be automated rather than hand coded for most lexical entries.

CO = consonantal

VO = vocalic

## GERMAN CONNECTIVE ADVERB FEATURES

CAT = CON

*CAN	canonical form
*ALO	allomorph
CU	conjunction use
SM	scope of modification
PO	position
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The connectives are those morphemes which may function as clause introducers or subordinating conjunctions as well as function within the clause. Typical examples include "allerdings", "deshalb", "ausserdem", etc. Grammatical rules will build a form entered as a CON to either a conjunction or to an argument or predicate modifier.

CU <conjunction use> indicates the function of a conjunction or a connective with respect to a following clause.

COR = coordinating  
SUB = subordinating  
INT = introductory

SM <scope of modification> indicates the range within the clause over which the connective's meaning may extend.

CLS = clause  
PRED = predicate  
ARG = any NP, PP, or CP used as a nominal argument  
ADJ = adjective

PO <position> specifies the potential location of the connective in relation to the components of the clause.

SI = sentence (clause) initial  
BEFVP = before verb phrase  
BEFMVB = before main verb  
SF = sentence final

## ENGLISH CONNECTIVE ADVERB FEATURES

CAT = CON

*CAN	canonical form
*ALO	allomorph
CU	conjunction use
SM	scope of modification
PO	position
ON	onset
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The connectives are those morphemes which may function as clause introducers or subordinating conjunctions as well as function within the clause. Typical examples include "however", "thus", "therefore", and "moreover". Grammatical rules will build a form entered as a CON to either a conjunction or to an argument or predicate modifier.

CU <conjunction use> indicates the function of a conjunction or a connective with respect to a following clause.

COR = coordinating (probably not needed for CON's)  
SUB = subordinating  
INT = introductory

SM <scope of modification> indicates the range within the clause over which the connective's meaning may extend.

CLS = clause  
PRED = predicate  
ARG = any NP, PP, or CP used as a nominal argument  
ADJ = adjective

PO <position> specifies the potential location of the connective in relation to the components of the clause.

SI = sentence (clause) initial  
BEFVP = before verb phrase  
BEFMVB = before main verb  
SF = sentence final

ON <onset> or quality of initial sound, is required for selection of "a" or "an". Although determiners do not typically precede a connective adverb, they can occur in this environment when a quotation, title, or semi-idiomatic expression follows.

CO = consonantal  
VO = vocalic

## GERMAN DEGREE ADVERB FEATURES

CAT = DEG

*CAN	canonical form
*ALO	allomorph
DG	degree (level of comparison)
SN	sign (direction of comparison)
PO	position
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The degree adverbs are a special category which function as adjective or adverb quantifiers. Included in this classification are the traditional degree markers such as "mehr" and "am meisten" as well as forms such as "ganz", "wirklich" and "sehr", which are sometimes considered intensifiers.

DG <degree> refers to the traditional levels of comparison.

POS = positive  
COM = comparative  
SUP = superlative

SN <sign> refers to the direction of comparison. Thus "am meisten" and "am wenigsten" are both superlative degree markers; however, they are opposite in the direction of comparison.

POS = positive ("mehr", "am meisten")  
NEG = negative ("weniger", "am wenigsten")

PO <position> indicates where the form lies with respect to other constituents within the sentence. Most of these have the obligatory value:

BEFADJ = before the adjective.

Several may also occur before another degree marker, for example "wirklich" and "ziemlich".

BEFDEG = before degree marker

## ENGLISH DEGREE ADVERB FEATURES

CAT = DEG

\*CAN canonical form  
\*ALO allomorph  
DG degree (level of comparison)  
SN sign (direction of comparison)  
PO position  
ON onset  
\*CAT lexical category  
\*PLC placement  
\*PRF preference  
\*LEX lexical collocation  
\*SNS sense number  
\*CNO concept number  
\*TAG area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The degree adverbs are a special category which function as adjective or adverb quantifiers. Included in this classification are the traditional degree markers such as "more" and "most" as well as forms such as "quite", "really", and "very", which are sometimes considered intensifiers.

DG <degree> refers to the traditional levels of comparison.

POS = positive  
COM = comparative  
SUP = superlative

SN <sign> refers to the direction of comparison. For example, "most" and "least" are both superlative degree markers; however, they are opposite in the direction of comparison.

POS = positive ("more", "most")  
NEG = negative ("less", "least")

PO <position> indicates where the form lies with respect to other constituents within the sentence. Most of these have the obligatory value:

BEFADJ = before the adjective.

Several may also occur before another degree marker, for example "really" and "somewhat".

BEFDEG = before degree marker

ON <onset> or quality of initial sound, is required for selection of "a" or "an".

CO = consonantal  
VO = vocalic

## GERMAN LOCATIVE ADVERB FEATURES

CAT = LOC

\*CAN canonical form  
\*ALO allomorph  
RO grammatical role  
PO position  
\*CAT lexical category  
\*PLC placement  
\*PRF preference  
\*LEX lexical collocation  
\*SNS sense number  
\*CNO concept number  
\*TAG area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The locative adverbials are those sentence modifiers which specify the locale of the activity. Such adverbials may be single word entries such as "da", phrases such as "weit und breit", or prepositional phrases such as "auf dem Tisch". Most prepositional phrase locatives are derived by the grammar from the individual words rather than being lexical entries themselves, but the single word entries and set phrases appear in the lexicon as terminal nodes.

RO <grammatical role> is the grammatical 'case' role function filled by the locative adverb. Any of the locative roles are possible, namely:

LOC = locative (subsumes STA and MOT)  
STA = station (subsumes PLC and PRX)  
PLC = place  
PRX = proximity  
MOT = motion (subsumes the following)  
CIR = circumferential  
ART = area traversed  
ORN = origin  
DST = destination

PO <position> represents potential positions within the sentence string  
a locative adverb may fill, typically:

AFTVP = after the VP  
SI = sentence initial

## ENGLISH LOCATIVE ADVERB FEATURES

CAT = LOC

*CAN	canonical form
*ALO	allomorph
RO	grammatical role
PO	position
ON	onset
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The locative adverbials are those sentence modifiers which specify the locale of the activity. Such adverbials may be single word entries such as "there", phrases such as "far and wide", or prepositional phrases such as "on the desk". Most prepositional phrase locatives are derived by the grammar from the individual words rather than being lexical entries themselves, but the single word entries and set phrases appear in the lexicon as terminal nodes.

RO <grammatical role> is the grammatical 'case' role function filled by the locative adverb. Any of the locative sets of roles are possible, namely:

LOC = locative (subsumes STA and MOT)  
STA = station (subsumes PLC and PRX)  
PLC = place  
PRX = proximity  
MOT = motion (subsumes the following)  
CIR = circumferential  
ART = area traversed  
ORN = origin  
DST = destination

PO <position> represents potential positions within the sentence string  
a locative adverb may fill, typically:

AFTVP = after the VP  
SI = sentence initial

ON <onset> or quality of initial sound, is required for selection of "a"  
or "an". Although determiners do not typically precede a locative  
adverb, they can occur in this environment when a quotation, title,  
or semi-idiomatic expression follows.

CO = consonantal  
VO = vocalic

## GERMAN MANNER ADVERB FEATURES

CAT = MAN

*CAN	canonical form
*AL0	allomorph
CL	morphological class
PO	position
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The manner adverbials are those which qualify the action of the verb by answering the question 'how'. Manner adverbials may be single word lexical entries such as "eilends" or "gerne" and are entered in the lexicon under this category. Manner adverbs which are derived from adjectives are not included as separate lexical entries.

CAN <canonical form> is the basic form of the manner adverb on which transfer operates. It represents all the possible allomorphic variants which the word may take and is used to group them together as a single semantic unit. Thus "gern", "lieber", and "am liebsten" would all have the same CAN.

AL0 <allomorph> is the actual surface string which is analyzed by the lexical entry. Each allomorphic variant thus has its own AL0 value.

CL <morphological class> indicates paradigmatic variation. In the instance of the manner adverbs, this is formation of the comparative and superlative. The values assigned are those which apply to the AL0 <allomorph>.

P-Ø = positive with no affix  
P-E = positive with an -e  
C-Ø = comparative with no affix  
C-ER = comparative with an -er  
S-Ø = superlative with no affix  
S-EST = superlative with an -est  
S-ST = superlative with an -st

PO <position> indicates potential positions taken by the manner adverb within the sentence string. The usual positions are:

BEFADJ = before adjective  
BEFMVB = before main verb  
AFTMVB = after main verb  
AFTOBJ = after object  
SI = sentence initial

## ENGLISH MANNER ADVERB FEATURES

CAT = MAN

*CAN	canonical form
*AL0	allomorph
CL	morphological class
PO	position
ON	onset
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The manner adverbials are those which qualify the action of the verb by answering the question 'how'. Most of the -LY adverbs fall into this category, although there are some exceptions. Also, certain forms which do not end in -LY are included, such as "well" and "fast". The manner adverbs which are derived from adjectives by the addition of -LY in a meaning-preserving derivation are not included as separate lexical entries. Rather, they are derived by the lexical rules from the adjective stem and appropriate affix as indicated by the adjective entry. Only those forms and senses which are not derivable are included as separate lexical entries.

CAN <canonical form> is the basic form of the manner adverb on which transfer operates. It represents all the possible allomorphic variants which the word may take and is used to group them together as a single semantic unit. Thus "well", "better", and "best" would all have the same CAN.

AL0 <allomorph> is the actual surface string which is covered by the lexical entry. Each allomorphic variant thus has its own AL0 value.

CL <morphological class> indicates paradigmatic variation. In the instance of the manner adverbs, this is formation of the comparative and superlative. The values assigned are those which apply to the ALO <allomorph>.

P-Ø = positive with no affix  
C-Ø = comparative with no affix  
C-M = comparative with "more"  
S-Ø = superlative with no affix  
S-M = superlative with "most"

PO <position> indicates potential positions taken by the manner adverb within the sentence string. The usual positions are:

BEFADJ = before adjective  
BEFMVB = before main verb  
AFTMVB = after main verb  
AFTOBJ = after object  
SI = sentence initial

ON <onset> or quality of initial sound, is required for selection of "a" or "an".

C0 = consonantal  
V0 = vocalic

## GERMAN ADVERBIAL PARTICLE FEATURES

CAT = PRT

*CAN	canonical form
*AL0	allomorph
PO	position
MD	mood
AS	aspect
SM	scope of modification
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The particles are those somewhat idiosyncratic forms which may reflect feelings about an utterance. Semantically their function is much like that of mood in verbs. Negatives and expletives are included in this category. Because the particles are so varied, their features are somewhat unique to each individual form; most, however, have the features described below. The following forms are likely to function as adverbial particles in German:

aber	gar
also	gerade
auch	halt
ausgerechnet	immer
bloss	ja
denn	naemlich
doch	nicht
eben	nicht (ein)mal
eigentlich	nun
einmal (mal)	nun (ein)mal
erst	nur
erst recht	schon
etwa	so
ganz	ueberhaupt
ganz und gar	vielleicht
	wohl

CAN <canonical form> is the basic form of the particle which is used in the transfer lexicon.

ALO <allomorph> is the actual surface string analyzed by the rule. Thus CAN (nicht) may have several ALO's including "nicht" and "un-".

PO <position> indicates the positions within the sentence in which the particle in question may occur. Particles are idiosyncratic in positioning and take any of a variety of positions, including:

SI = sentence initial  
SF = sentence final  
BEFMVB = before main verb  
BEFNP = before NP

MD <mood> is the same as mood associated with verbs. The particles can be used to indicate mood where it is not inflectionally marked on the verb.

IND = indicative  
SUB = subjunctive  
IRR = irrealis (contrary to fact)

AS <aspect> is also a verb feature which is sometimes carried by means of a particle rather than verbal inflection.

PER = perfective  
IMP = imperfective

SM <scope of modification> is used to indicate the level at which the particle functions, e.g., whether it refers to the entire sentence, just the verb phrase, etc. The values are potentially any non-terminal node, but typically the following seem to occur:

S = entire sentence  
VP = verb phrase  
NP = noun phrase  
ADV = adverb

## ENGLISH ADVERBIAL PARTICLE FEATURES

CAT = PRT

*CAN	canonical form
*ALO	allomorph
PO	position
MD	mood
AS	aspect
SM	scope of modification
ON	onset
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The particles are those somewhat idiosyncratic forms which may reflect feelings about an utterance. Semantically their function is much like that of mood in verbs. Negatives and expletives are included in this category. Because the particles are so varied, their features are somewhat unique to each individual form; most, however, have the features described below.

ALO <allomorph> is the actual surface string covered by the rule. Thus CAN (NOT) has several ALO's including NOT, N'T (e.g., "doesn't"), and N (e.g., "never").

PO <position> indicates the positions within the sentence in which the particle in question may occur. Particles are idiosyncratic in positioning and take any of a variety of positions, including:

SI	= sentence initial
SF	= sentence final
BEMV	= before main verb
BEFNP	= before NP

MD <mood> is the same as mood associated with verbs. The particles can be used to indicate mood where it is not inflectionally marked on the verb.

IND = indicative  
SUB = subjunctive  
IRR = irrealis (contrary to fact)

AS <aspect> is also a verb feature which is sometimes carried by means of a particle rather than verbal inflection.

PER = perfective  
IMP = imperfective

SM <scope of modification> is used to indicate the level at which the particle functions, e.g., whether it refers to the entire sentence, just the verb phrase, etc. The values are potentially any non-terminal node, but typically the following seem to occur:

S = entire sentence  
VP = verb phrase  
NP = noun phrase  
ADV = adverb

ON <onset> or quality of initial sound, is required for selection of "a" or "an". Although determiners do not typically precede an adverbial particle, they can occur in this environment when a quotation, title, or semi-idiomatic expression follows.

CO = consonantal  
VO = vocalic

## GERMAN PREFIX FEATURES

### CAT = PRFX

*CAN	canonical form
*AL0	allomorph
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

This category is primarily for those elements which may be prefixed to German verbs. Since the co-occurrence restrictions between prefixes and verbs are usually included in the verb description (i.e., each verb lists its possible prefixes in its features), only the minimum array of values is required.

## ENGLISH ADPREP ('PREFIX') FEATURES

CAT = PRFX

Many English verbs take an adverbial particle, homographic with a preposition, to complete their meaning. They are sometimes referred to as 'phrasal verbs'. That the particle is not a true 'preposition' may be illustrated by the fact that a following noun is not always required as an object, as it would be for a preposition, e.g.:

The car broke DOWN  
The excitement died AWAY  
The teacher showed UP  
The shoes wore OUT

Although they appear only postposed rather than prefixed to the verb in English, such forms are somewhat similar to German separable prefixes in terms of function. Thus, for the sake of convenience, the constituent is given the label PRFX in English as well. Another term sometimes seen for these is 'adprep'.

When the verb in question is transitive, the particle is often placed following the object rather than immediately following the verb, e.g.,

He bought his partner out  
He bought out his partner

When the object is a pronoun, this postposed position is obligatory

He bought him out  
\*He bought out him

This test is useful in distinguishing between such particles and true prepositional objects, as in

He appealed to her  
He dabbles in it  
They stumbled across it  
She worried about him

These true prepositional objects are indicated under the MA <syntactic marker of argument> feature of the verb and are treated as prepositions in the system.

Note that a verb may sometimes take the same 'preposition' as the marker of a prepositional object (MA) and as an adprep (PX):

She turned him on	Adprep 'prefix'
She turned on him	Prepositional object
or	
They tore up the street (with air-hammers).	PX
They tore up the street (as fast as they could go).	MA

Potentially ambiguous combinations such as the above often restrict positioning of the prefix such that it is only postposed to the noun phrase rather than allowing the preposed position as well. This is accounted for under the PM <prefix mobility> feature of the verb.

At present the adverbial/prepositional particles are given no special features other than the usual ones described at the beginning of this appendix:

*CAN	canonical form
*ALO	allomorph
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

## GERMAN PREVERB FEATURES

CAT = PREVB

\*CAN canonical form  
\*AL0 allomorph  
PO position  
MD mood  
\*CAT lexical category  
\*PLC placement  
\*PRF preference  
\*LEX lexical collocation  
\*SNS sense number  
\*CNO concept number  
\*TAG area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The preverbs are those adverbs which are characteristically associated with the verb but which are not manner adverbials. They tend to qualify the activity set forth in the verb phrase. Typical examples include "immer", "selten", "fast", "ungewoehnlich". Forms such as "vielleicht" and "hoffentlich" can also be included in this class.

PO <position> indicates potential locations of the preverb within the sentence string. The most likely position is before the main verb (and after the modal, if present), although occasionally other positions may be possible.

BEFMVB = before main verb  
BEFVP = before verb phrase  
SI = sentence initial

MD <mood> is identical to the traditional concept of mood as is associated with verb paradigms, and has the same values.

IND = indicative  
SUB = subjunctive  
IRR = irrealis (contrary to fact)

## ENGLISH PREVERB FEATURES

CAT = PREVB

\*CAN canonical form  
\*ALO allomorph  
PO position  
MD mood  
ON onset  
\*CAT lexical category  
\*PLC placement  
\*PRF preference  
\*LEX lexical collocation  
\*SNS sense number  
\*CNO concept number  
\*TAG area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The preverbs are those adverbs which are characteristically associated with the verb but which are not manner adverbials. They tend to qualify the activity set forth in the verb phrase. Typical examples include "always", "seldom", "almost" "rarely". Forms such as "perhaps" and "hopefully" can also be included in this class.

PO <position> indicates potential locations of the preverb within the sentence string. The most likely position is before the main verb (and after the modal, if present), although occasionally other positions may be possible.

REFMVB = before main verb  
BEFVP = before verb phrase  
SI = sentence initial

MD <mood> is identical to the traditional concept of mood as is associated with verb paradigms. Languages which do not routinely inflect for mood (as is often true of English) may mark mood by means of preverbs. The values are the same as MD for verbs.

IND = indicative  
SUB = subjunctive  
IRR = irrealis (contrary to fact)

ON <onset> or quality of initial sound, is required for selection of "a"  
or "an".

CO = consonantal

VO = vocalic

## GERMAN TEMPORAL ADVERB FEATURES

CAT = TMP

*CAN	canonical form
*ALO	allomorph
RO	grammatical role
PO	position
AS	aspect
TN	tense
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The temporal adverbials provide information about when the activity or state described by the sentence takes place or is in effect. They frequently answer the question "when" and sometimes impart aspectual information. Temporal adverbials may be single word lexical entries such as "jetzt" or "morgen", and are entered in the lexicon under this category. However, they may also be derived by the syntactic rules as phrases containing a head noun with semantic type values indicating temporality, i.e., TY (TMP). Examples include "naechste Woche", "letztes Jahr", "vor einigen Monaten", and "am Morgen".

RO <role> is grammatical 'case' role as proposed in Fillmore's case grammar. Temporal adverbs will have either the general value TMP (temporal) or one of the more specific values: DUR (durative), e.g., "wahrend des Konzerts", or PCT (punctual), e.g., "um 3 Uhr".

PO <position> indicates potential locations of the temporal adverb within the sentence. Usual values are:

SF = sentence final  
SI = sentence initial

AS <aspect> values are identical to those of aspect with respect to verbs.

PER = perfective  
IMP = imperfective

TN <grammatical tense> specifies temporal relevance.

PR = present  
PA = past  
FU = future

## ENGLISH TEMPORAL ADVERB FEATURES

CAT = TMP

*CAN	canonical form
*ALG	allomorph
RO	grammatical role
PO	position
AS	aspect
TN	tense
CN	onset
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The temporal adverbials provide information about when the activity or state described by the sentence takes place or is in effect. They frequently answer the question "when" and sometimes impart aspectual information. Temporal adverbials may be single word lexical entries such as "now" or "tomorrow", in which case they are entered in the lexicon under this category. However, they may also be derived by the syntactic rules as phrases containing a head noun with semantic type values indicating temporality, i.e., TY (TMP). Examples include "next week", "last year", "months ago", and "in the morning".

PO <role> is grammatical case role as proposed in Fillmore's case grammar. Temporal adverbs will have either the general value TMP (temporal) or one of the more specific values: DUR (durative), e.g., "during the concert", or PCT (punctual), e.g., "at 3 o'clock".

PO <position> indicates potential locations of the temporal adverb within the sentence. Usual values are:

SF = sentence final  
SI = sentence initial

AS <aspect> values are identical to those of aspect with respect to verbs.

PER = perfective  
IMP = imperfective

TN <grammatical tense> specifies temporal relevance.

PR = present  
PA = past  
FU = future

ON <onset> or quality of initial sound, is required for selection of "a" or "an". Although determiners do not typically precede a temporal adverb, they can occur in this environment when a quotation, title, or semi-idiomatic expression follows.

CO = consonantal  
VO = vocalic

## GERMAN CONJUNCTION FEATURES

CAT = CONJ

*CAN	canonical form
*ALO	allomorph
CU	use of conjunction
CJ	form of conjoined clause
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

CU <conjunction use> is used to mark the function of the conjunction in conjoining phrases or clauses.

COR = coordinating  
SUB = subordinating  
INT = introductory

CJ <form of conjoined clause> is used to mark the types of phrases or clauses which can be conjoined by the conjunction in question.

COR = coordinate clause  
SUB = subordinate clause  
INF = infinitive phrase  
MCL = main clause

## ENGLISH CONJUNCTION FEATURES

CAT = CONJ

*CAN	canonical form
*ALO	allelomorph
CU	use of conjunction
CJ	form of conjoined clause
ON	onset quality
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

CU <conjunction use> specifies the kind of connective function.

COR = coordinating  
SUB = subordinating  
INT = introductory

CJ <form of conjoined clause> marks the types of phrases or clauses which can be conjoined by the conjunction in question.

COR = coordinate clause  
SUB = subordinate clause  
INF = infinitive phrase  
MCL = main clause

ON <onset> or quality of initial sound, is required for selection of "a" or "an". Although determiners do not typically precede a conjunction, they can occur in this environment when a quotation, title, or semi-idiomatic expression follows.

CO = consonantal  
VO = vocalic

## GERMAN DETERMINER FEATURES

CAT = DET and DET  
/ \  
DST D-FLEX

For parsing purposes German determiners have been divided into two classes: DST (determiner stem) and DET (determiner).

The DST class consists of determiners which will be analyzed by the system at the lexical level as stem plus ending and then combined at the word level to form a DET. Almost all German determiners are derived in this way. Examples: "keiner", "eir", "der", "diese".

The DET class is a much smaller group of determiners whose complex or deviant structure does not make it feasible to derive them in this manner. Combined determiners of the form "dasjenige" or "derselbe" and the enclitic determiners in prepositions like "ans", "beim", or "zur" fall into this category. These determiners will be recognized as DET at the lexical level and will not be further analyzed at the word level. Thus DET is both a lexical and a word class, while DST is a lexical class only:

analyzed at	:	:	
word level as	DET	:	
	/ \	:	
analyzed at		:	
lexical level as	DST D-FLEX	DET	
	/ \ \		
parsed as	kein	er	dasjenige
input string	keiner		dasjenige

Some of the features discussed below are shared in common by DST and DET, while others apply exclusively to one class or the other. Each feature has been appropriately marked in the list below to avoid confusion.

The coding features for German determiners are:

*CAN	canonical form	- BOTH
*ALO	allomorph	- BOTH
CL	inflectional class	- DST
GD	grammatical gender	- DET
CA	grammatical case	- DET
NU	grammatical number	- DET
IN	inflection	- DET
KD	kind of determiner	- BOTH
NR	number restriction	- BOTH
BF	bound form	- DET
SY	syncopated letter	- DST
SX	sex (natural gender)	- DST
*CAT	lexical category	- BOTH
*PLC	placement	- BOTH
*PRF	preference	- BOTH
*LEX	lexical collocation	- BOTH
*SNS	sense number	- BOTH
*CNO	concept number	- BOTH
*TAG	area of provenience tag	- BOTH

(Asterisks mark systemic features described generally at the beginning of this appendix)

CAN <canonical form> The value for CAN is the nominative singular neuter form of the lexical entry or, in instances where the determiner is restricted to the plural, it is the nominative plural form. For the enclitic determiners "m", "n", "r" and "s", the value for CAN is the nominative singular of the unbound form.

ALO <allomorph> has as its value the lexical string itself. Where CAN <canonical form> and ALO have the same value, ALO may be omitted in coding.

CL <morphological class> is used only with members of DST. Its values are typical determiners, each of which represents an inflection pattern. This feature enables the system to limit the range of possible genders, cases, and numbers when it identifies a DST combined with an appropriate suffix.

EXAMPLE		S I N G U L A R				P L U R A L			
		masculine	feminine	neuter		m,f,n			
CL	stem	N G D A	N G D A	N G D A	N G D A				
DER	d-	er es em en	ie er er ie	as es em as	ie er en ie				
SEIN	sein-	Ø es em en	e er er e	Ø es em Ø	e er en e				
DIESER	dies-	er es em en	e er er e	es es em es	e er en e				
ALLER	all-	er en em en	e er er e	es en em es	e er en e				
WELCHER	welch-	er es em en	e er er e	es es em es	e er en e				
		en		en					

CD <grammatical gender> is a feature of the category DET.

M = masculine  
 F = feminine  
 N = neuter

CA <grammatical case> is also a German DET feature.

N = nominative  
 G = genitive  
 D = dative  
 A = accusative

NU <grammatical number> is another DET feature.

SG = singular  
 PL = plural

IN <inflection> is also a DET feature. Most determiners are strongly inflected. For these, the subsequent adjectives must be weakly inflected. However, there are some determiners which show weak inflection and require strong inflection of the adjectives which follow.

ST = strong inflection ("der hohe Turm")  
 WK = weak inflection ("ein hoher Turm")

KD <kind of determiner> In German (and in English as well), the determiner and pronoun classes share a number of members. Lexical entries that can belong to either class are listed in this system as determiners and marked with the feature KD. Values for KD include the subclass or subclasses of pronoun to which the word can also belong.

DET = determiner  
DEM = demonstrative pronoun  
REL = relative pronoun  
IND = indefinite pronoun  
INT = interrogative pronoun  
POS = possessive pronoun

NR <number restriction> is coded only for those determiners (DST or DET) whose use is restricted with regard to number. Some determiners occur in the singular but not in the plural. Others occur in the plural but not in the singular. And a few determiners can occur unrestricted with plural noun phrases but do not occur with singular noun phrases unless the determined common noun is a mass noun.

SG = singular only ("ein": ein Mann, eine Frau)

PL = plural only ("einige": einige Leute)

MS PL = singular mass noun and plural only  
("alles": alles Bier, alle Kinder)

BF <bound form> is used for determiners (pronouns) which may occur as the second element in a contracted form with a preposition. The value of BF is the string itself, for example M (as in "im") or R (as in "zur").

SY <syncopated> is used in those determiners which are syncopated forms, such as "unsr-". The value given SY is the deleted letter.

SX <natural gender> is used for determiners such as "sein", etc., which may have nouns with natural gender qualities as referents yet may modify a noun with any grammatical gender. Whereas grammatical gender is not used in the transfer process natural gender is.

M = male referent  
F = female referent  
N = neuter referent

## ENGLISH DETERMINER FEATURES

CAT = DET

*CAN	canonical form
*ALO	allomorph
NU	grammatical number
NR	number restriction
KD	kind of determiner
SX	sex (natural gender)
ON	onset required of following word
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: English determiners include all forms which occur as prenominal determiners, e.g., "a", "the", "that", "some", "her".

CAN <canonical form> is the basic form of the determiner. The form which would be used modifying a singular subject is to be preferred (nominative singular), but where no singular exists, the plural is used.

ALO <allomorph> is the actual surface form of the string. When the CAN <canonical form> and ALO are identical, ALO may be omitted in coding.

NU <grammatical number> is as one would expect:

SG = singular  
PL = plural

NR <number restriction> is coded only for those determiners whose use is restricted with regard to number. Some determiners occur in the singular but not in the plural. Others occur in the plural but not in the singular. And a few determiners can occur unrestricted with plural noun phrases but do not occur with singular noun phrases unless the determined common noun is a mass noun.

SG = singular only ("a"/"an")  
PL = plural only ("several")  
MA PL = singular mass noun and plural only ("all")

KD <kind of determiner> In English (and in German as well), the determiner and pronoun classes share a number of members. Lexical entries that can belong to either class are listed in this system as determiners and marked with the feature KD. Values for KD include the subclass or subclasses of pronoun to which the word can also belong.

DET = determiner  
DEM = demonstrative pronoun  
REL = relative pronoun  
IND = indefinite pronoun  
POS = possessive pronoun  
INT = interrogative pronoun

SX <natural gender> is marked for determiners which have inherent sexuality, such as "her" or "his".

M = male  
F = female  
N = neuter or unmarked

ON <onset> or quality of initial sound, is required for selection of "a" or "an". Although determiners do not typically precede another determiner, they can occur in this environment when a quotation, title, or semi-idiomatic expression follows.

CO = consonantal  
VO = vocalic

## GERMAN PREDETERMINER FEATURES

CAT = PREDET

*CAN	canonical form
*ALO	allomorph
NU	grammatical number
NR	semantic number restriction
DR	determiner requirement
PO	position
SQ	sequence
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The predeterminers are those morphemes which may occur preposed to the usual determiners. They share properties with DET's in that they may occur in the NP preposed to a noun. Some, like adjectives, can be used following the copula. Typical examples are "all", "solch", and "nur". Also included as PREDET's are a number of phrasal forms which are derived by the grammars. Most notable are phrases consisting of a determiner or a noun of quantity together with another noun. Typical examples are "ein Liter (Milch)", "ein Kilogramm (Kartoffeln)", "ein Dutzend (Eier)".

NU <grammatical number> refers to the number required of the following noun for agreement purposes.

SG = singular  
PL = plural

NR <number restriction> is used for semantic number. Certain nouns such as mass nouns and collectives may be singular in form but accept certain PREDET's which usually require plural nouns, for example "all" and "solch". These PREDET's are identified by NR as allowing semantic number to take precedence over syntactic number. The relevant value is:

PLDT = plural determiner

DR <determiner requirement> indicates whether the following nominal can be a noun which stands alone as an NP or whether there must be a determiner between the PREDET and the NO. Note that the feature may be linked to the NU <grammatical number> and NR <number restriction> values. For example, "all" requires a determiner if it modifies a singular or mass noun, but if "all" modifies a plural form, the determiner is optional.

RD = requires a determiner

NP = any NP

PO <position> specifies the sentential positions where the PREDET may be found. The usual values are:

BEFMOD = before modificand

AFTCOP = after copula

AFTMOD = after modificand

SQ <sequence> indicates the location of the PREDET within a string of prenominal modifiers. The values indicate the slot within which the PREDET falls. The relevant value is:

PDT = predeterminer position

## ENGLISH PREDETERMINER FEATURES

CAT = PREDET

*CAN	canonical form
*ALO	allomorph
NU	grammatical number
NR	semantic number restriction
DR	determiner requirement
PO	position
SQ	sequence
ON	onset
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: The predeterminers are those morphemes which may occur preposed to the usual determiners. They share properties with DET's in that they may occur in the NP preposed to a noun. Some, like adjectives, can be used following the copula. Typical examples are "all", "such", and "only". Also included as PREDET's are a number of phrasal forms which are derived by the grammars. Most notable are phrases consisting of a determiner or a noun of quantity together with the preposition "of". Typical examples are "most of", "some of", "a quart of", "a mile of", "two dozen of".

NU (grammatical number) refers to the number required of the following for agreement purposes.

SG = singular  
PL = plural

NR <number restriction> is used for semantic number. Certain nouns such as mass nouns and collectives may be singular in form but accept certain PREDET's which usually require plural nouns, for example "all", "some", "each of". These PREDET's are identified by NR as allowing semantic number to take precedence over syntactic number. The relevant value is:

PLDT = plural determiner

DR <determiner requirement> indicates whether the following nominal can be a noun which stands alone as an NP or whether there must be a determiner between the PREDET and the NO. Thus some PREDET's such as "some of" require a determiner, whereas the determiner is optional with "all". Note that the feature may be linked to the NU <grammatical number> and NR <number restriction> values. For example, "all" requires a determiner if it modifies a singular or mass noun, but if it modifies a plural form, the determiner is optional.

RD = requires a determiner  
NP = any NP

PO <position> specifies the sentential positions where the PREDET may be found. The usual values are:

BEFMOD = before modificand  
AFTCOP = after copula  
AFTMOD = after modificand

SQ <sequence> indicates the location of the PREDET within a string of prenominal modifiers. The values indicate the slot within which the PREDET falls. The relevant value is:

PDT = predeterminer position

ON <onset> or quality of initial sound, is required for selection of "a" or "an". Although determiners do not typically precede a predeterminer, they can occur in this environment when a quotation, title, or semi-idiomatic expression follows.

CO = consonantal  
VO = vocalic

## GERMAN NOUN FEATURES

CAT = NST

*CAN	canonical form
*ALO	allomorph
CL	inflectional class
{CA}	grammatical case
{NU}	grammatical number
NR	number restriction
GD	grammatical gender
CP	derivatives capitalized (proper name)
DR	determiner requirement
TY	semantic type
SX	sex (natural gender)
RC	'case' role of complement
TC	semantic type of complement
NC	surface marker of complement
FC	syntactic form of complement
HG	homographic with
DF	derivable from
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix. Those set off by braces are not coded but are brought in by the CL <class> values. They are carried on the N-FLEX <noun inflectional ending> and are introduced in the morphological phase.)

CAN <canonical form> A quasi-canonical form of the word used to identify lexical entries. For nouns, it is the nominative singular form if one exists. If there is no singular, the nominative plural form is used. When a word has multiple senses which require different configurations of values, it is accompanied by a uniquely identifying sense number (e.g., "TOR 1" for the meaning "gate"; "TOR 2" for the meaning "goal"; and "TOR 3" for the meaning "fool"). Each sense (or group of senses with the same configuration of values) is coded as a separate entry.

ALO <allomorph> is the stem on which a particular configuration of inflectional endings specified by a CL <class> value may attach. The examples shown under CL illustrate the various kinds which may occur.

CL <inflectional class> values correspond to the particular configuration of morphological affixes possible with the ALO <allomorph>. One set of values designates a combination of nominative, genitive, dative, and accusative affixes in the singular. The second set represents a combination of those affixes in the plural.

### S I N G U L A R

Nomin- ative	Geni- tive	Dative	Accu- sative	Symbol	Example stems
Ø	Ø	Ø	Ø	S-A	Tafel, Hand, Kenntnis
Ø	-s	Ø	Ø	S-B	Garten, Maurer, Clown
Ø	-ses	Ø	Ø	S-C	Atlas, Buendnis
Ø	-es	Ø	Ø	S-D	Haus, Kreuz, Schmerz
Ø	-es/-s	Ø	Ø	S-E	Ball, Buch, Gestuehl
Ø	-n	-n	Ø	S-F	die Illustrierte
Ø	-ens	-en	Ø	S-G	Herz
Ø	-s/-n	Ø/-n	Ø/-n	S-H	Nachbar
Ø	-n	-n	-n	S-I	Genosse, Junge, Herr
Ø	-ns	-n	-n	S-J	Wille, Name, Gedanke
Ø	-en	-en	-en	S-K	Zar, Baer, Student
-us	-us	-us	-us	S-L	Typ-(us), Isthm-(us)
Ø/-n	-ns/-n	-n	-n	S-M	Friede
-r	-n	-n	-n	S-N	Abgeordnete-
-um	-ums	-um	-um	S-O	Atri-(um), Muse-(um)
-s	-n	-n	-s	S-P	ein Geraeucherte-
Ø	-es/-s	Ø/-e	Ø	S-Q	Fall

### P L U R A L

Nomin- ative	Geni- tive	Dative	Accu- sative	Symbol	Example stems
Ø	Ø	Ø	Ø	P-1	Gaerten, Marginalien, Bestreben
Ø	Ø	-n	Ø	P-2	Leute, Maurer, Gebirge, Voegel
-e	-e	-en	-e	P-3	Gestuehl, Kreuz, Haend, Faell
-er	-er	-ern	-er	P-4	Buech, Geld, Haeus, Maenn
-en	-en	-en	-en	P-5	Schmerz, Tat, Atri-(um), Zar
-n	-n	-n	-n	P-6	Nachbar, Tafel, Birne, Auge
-nen	-nen	-nen	-nen	P-7	Abenteuerin, Embryo
-s	-s	-s	-s	P-8	Alibi, Clown, Bar
-se	-se	-sen	-se	P-9	Atlas, Buendnis, Kenntnis
-e	-er	-en	-e	P-10	Abgeordnet-, Illustriert- (but without preceding determiner)

CA <grammatical case>

N = nominative  
G = genitive  
D = dative  
A = accusative

NU <grammatical number>

SG = singular  
PL = plural

MR <number restriction> marks nouns whose numerical character is not that of simple singular or plural.

PLVB = singular form of the noun may take a plural verb  
PLDT = singular form of noun may take a plural determiner

GD <grammatical gender> is a normal morphological feature of German and is distinct from natural gender, which is covered under SX <sex> where relevant. Grammatical gender is important in identifying relationships between nouns and potential noun modifiers in German. It often provides an important clue in determining syntactic structure, since nouns and their modifiers must agree with respect to gender.

M = masculine  
F = feminine  
N = neuter

CP <capitalization> is used to indicate forms which would traditionally be considered "proper nouns". They are usually names of persons or places or the trademark designation of a product. This needs to be coded for German, which capitalizes all nouns, because derivatives will also frequently be capitalized, as will their equivalents in other European languages.

UC = ("upper case") form is a proper noun  
LC = ("lower case") form is not a proper noun [DEFAULT]

DR <determiner requirement> specifies the syntactic character of a given noun. Some nouns such as "Deutschland" or "Hans" may stand alone as noun phrases without an accompanying determiner, while other nouns such as "Schweiz" or "Mittelmeer" cannot. Likewise, most mass nouns can stand alone without a determiner.

NP = occurs only without a determiner

RD = requires a determiner

RD NP = may occur with or without a determiner (i.e. determiner is translated if it is present in the source language)

TY <semantic type> is used to characterize a broad semantic class of nouns with which a particular noun may be associated. Such distinctions are frequently useful in disambiguation. For example, the noun "Hahn" has several meanings in German. However, if it is used with a verb such as "sterben" (der Hahn stirbt), "Hahn" may be readily disambiguated (= "rooster" or "cock"), since the verb "sterben" requires an animate subject. If it is used with a verb like "entzweigehen" (der Hahn geht entzwei), "Hahn" may again be disambiguated (= "faucet" or "valve") since this particular verb takes an inanimate subject. The values for semantic type are represented in the form of a matrix, presented here in Appendix A-2.

SX <sex - natural gender> is used for nouns which have inherent sexuality, primarily animates. Note that this value does not always agree with grammatical gender.

M = male  
F = female  
N = neuter [DEFAULT]

RC <role of complement> The relationship of the complement in regard to the noun is specified as a role similar to those introduced in the 'case grammar' approach to linguistic theory. The value is an acronym from the list of such roles given in Appendix A-3. Common examples include:

ORN = origin or source  
DST = destination of goal  
PUR = purpose or reason  
BEN = benefactive  
MAN = manner

TC <semantic type of complement> is used to designate any semantic restrictions upon the noun which may function as a complement to the one the entry describes. The values are the same as for TY <semantic type> above.

MC <surface marker of complement> The complement of a noun may be marked by a particular case or by a preposition, and the values are the acronym for the case or the preposition(s) which fulfill this function. All nouns have a default value supplied by the system to permit genitive complements, so this case value is not coded.

G = genitive [DEFAULT]  
A = accusative  
— = (any prepositions, the value being their canonical form)  
TH = "dass" complement ("die Aussage, dass...")  
FT = 'infinitive' complement ('der Drang zu trinken')  
WH = subordinate question as complement

FC <form of complement> specifies the kind(s) of syntactic constituent which may function as the complement of the noun.

CP = complement phrase  
NP = noun phrase  
PP = prepositional phrase

HG <homograph> marks nouns which have the same spelling as another part of speech. The values indicate the syntactic class of the form matched.

ADJ = adjective  
ADV = adverb  
INF = infinitive  
OTR = other

Although this feature is not used by the system at present, it may be included in the current lexical entries as the basis for disambiguation rules to be incorporated into the finished system.

DF <derivable from> is assigned to derived nominals (usually deverbal forms such as gerunds and agentive nouns, but also certain nouns derived from adjective stems or other nouns). It particularizes the underlying word in the event that some of its grammatical characteristics may have been carried along during derivation.

VI = intransitive verb  
VT = transitive verb  
VR = reflexive verb  
ADJ = adjective  
NO = another noun

## ENGLISH NOUN FEATURES

CAT = NST

*CAN	canonical form
*ALO	allomorph
CL	inflectional class
{CA}	grammatical case
{NU}	grammatical number
NR	number restriction
CP	capitalization
DR	determiner requirement
TY	semantic type
SX	sex (natural gender)
RC	'case' role of complement
TC	semantic type of complement
MC	surface marker of complement
FC	syntactic form of complement
HG	homographic with
DF	derivable from
ON	onset (quality of initial sound)
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix. Those set off by {} are brought in by means of the CL <inflectional class> feature during processing rather than coded separately.)

CAN A quasi-canonical form of the word, it is its nominative singular if one exists. (If it does not, as for instance "trousers", the nominative plural is used.) When the word has multiple senses each different sense of the noun is coded separately, with its CAN accompanied by an identificatory sense number. Like the SNS <sense number> feature, this may match the definition number of the word in one of the dictionaries used for reference, but unlike the SNS value it will not be preceded by a code letter for the dictionary.

ALO <allomorph> In regular nouns which form their plurals and possessives by addition of suffixes to a common stem, the ALO is the sequence of letters which form that stem. For such nouns, there is only a single ALO for each CAN <canonical form>, namely, the sequence of letters which remains the same whether the word is singular or plural, nominative or possessive. For example, the upper case portions of the following words:

CODex	ACTIVITY	COW	SHEEP	RADIus
CODex'	ACTIVITY's	COW's	SHEEP's	RADIus'
CODices	ACTIVITIES	COWs	SHEEP	RADIi
CODices'	ACTIVITIES'	COWs'	SHEEP's	RADIi's

The exceptions are the irregular nouns such as "man/men", "mouse/mice", or "leaf/leaves", which require different ALO's for singular and plural, and hence two lexical entries per sense. Any ALO which is identical with its CAN <canonical form> will be generated automatically.

If a sense occurs only as a plural, e.g., "minutes" in the sense of 'record of proceedings', that plural is coded as the ALO, with the CAN <canonical form> in the singular. If a word occurs only as a plural, e.g., "scissors", that plural is coded as the CAN, and generated later as the ALO). The examples shown under CL <inflectional class> illustrate varieties of ALO's.

CL <morphological class> takes values which correspond to the singular and plural, nominative and possessive affixes pertinent to the ALO <allomorph>. One value represents a pair of nominative and possessive affixes in the singular. The second designates a pair of nominative and possessive plural affixes. A noun has a zero ( $\emptyset$ ) affix when it is the exact same string of letters as the ALO <allomorph> (see COW and SHEEP, nominative singular, and SHEEP, nominative plural, above). When a noun simply does not occur in a given case/number, no value corresponding to them is coded. "Celibacy", for instance, is never plural, so it has no value for nominative and possessive plural.

The charts which follow illustrate the system for deriving the two values for CL <inflectional class>. (The hyphen after the stem, used there to indicate that a word is truncated, is not coded.)

### S I N G U L A R

Nomin- ative	Posses- sive	Symbol	Example stems
Ø	'S	S-Ø1	work, sheep, child, seraph, stroma, A, b, b, mouse, wife, man, flambeau, alumna, aircraft, celibacy, case, monsignor, femur, approach, wish
Ø	'	S-Ø2	apparatus, optics, series, sphinx
Y	Y'S	S-Y	stud-, intensit-
IS	IS'	S-IS	cris-, analys-, ephemera-
US	US'	S-US	radi-, alumna-, gen-, op-, corp-
X	X'	S-X	laryn-, matri-
EX	EX'	S-EX	ind-, cod-
ON	ON'S	S-ON	criteri-, automat-
UM	UM'S	S-UM	dat-, curricul-
n o v a l u e c o d e d			
			children, mice, wives, femora, men, scissors, [business] affairs

### P L U R A L

Nomin- ative	Posses- sive	Symbol	Example stems
S	S'	P-S	work, case
ES	ES'	P-ES	apparatus, sphinx, cris-, analys-, approach, wish
IES	IES'	P-IES	stud-, intensit-
Ø	'S	P-Ø1	men, children, sheep, aircraft, mice ,
Ø	'	P-Ø2	series, scissors, trousers, wives, [business] affairs, femora
A	A'S	P-A	criteri-, automat-, dat-, curricul-
TA	TA'S	P-TA	stroma
E	E'S	P-E	stria, alumna
I	I'S	P-I	radi-, alumna-, monsignor-
IM	IM'S	P-IM	seraph
X	X'	P-X	flambeau
CES	CES'	P-CES	matri-
GES	GES'	P-GES	laryn-
'S	'S	P-ØS	A, b, 6
IDES	IDES'	P-IDES	ephemer-
ICES	ICES'	P-ICES	ind-, cod-
ERA	S	P-ERA	gen-, op-
ORA	ORA'S	P-ORA	corp-
n o v a l u e c o d e d			
			celibacy, child, mouse, wife, man, femur

Because the use of the "'s" with words ending in /s/ or /z/  
sounds is spreading, we have compromised by using that form  
of the possessive affix on any such words which end in "-e".

CA <grammatical case> is coded in the lexicon on the N-FLEX entries rather than the noun stems. It is introduced by the grammars on the basis of CL <inflectional class> values coming from the input text, and indicates potential sentential functions similar to case restrictions in German.

U = unmarked case (can be subject or object)  
P = possessive

NU <grammatical number> is also a feature of the N-FLEX category rather than the stem. During the morphological stage of processing it is brought in by the grammars on the basis of the input CL <inflectional class> values and is used for verb agreement.

SG = singular  
PL = plural

NR <number restriction> marks nouns whose numerical character is not that of simple singular/plural.

PLVB = singular form may take a plural verb  
PLDT = singular form may take a plural determiner

CP <capitalized> When a noun in a given sense is always capitalized, the value UC is given. When it is likely to be capitalized in circumstances other than at the beginning of a sentence, as a personification, or in a book title, UC LC is marked. A form that is typically not capitalized is given LC only.

UC = always capitalized ("April", "Mississippi", "Dr.",  
[the Judaeo-Christian] "God", "St.")

UC LC = sometimes capitalized ("the tower/Chrysler tower",  
"our street/Main Street", "the ballet/Ballet Russe  
de Monte Carlo", "doctor/Doctor")

LC = usually lower case only [DEFAULT]

DR <determiner requirement> concerns the syntactic character of certain nouns with regard to determiners. Some, such as "Switzerland" or "Mars", may stand alone as noun phrases without an accompanying determiner, while other nouns such as "the Netherlands" cannot.

NP = occurs only without a determiner

RD = requires a determiner

RD NP = may occur with or without a determiner (i.e., determiner is translated if it is present in the source language)

TY <semantic type> is used to enable semantic co-occurrence restrictions with modifiers and verbs to be established according to a matrix of plus and minus values. See Appendix A-2 for the list of values.

SX <sex - natural gender> is indicated for nouns which have inherent sexuality or which are characteristically referred to by pronouns usually used for male and female animates.

M = male (masculine pronoun)

F = female (feminine pronoun)

N = unmarked (neuter pronoun) - [DEFAULT]

RC <role of complement> The semantic relationship between the complement and the modified noun is described in terms of a 'case grammar' role. The values are the same as those used with the verbs and are listed in Appendix A-3. Examples include:

ORN = origin or source ("transportation from")

DST = destination or goal ("transportation to")

PUR = purpose or reason

BEN = benefactive

MAN = manner

TC <type of complement> Semantic type must also be specified for a nominal which functions as a noun complement. The values, like those for TY <semantic type> above, are listed in Appendix A-2.

MC <complement marker> Although at present the larger array of complements required for German nouns does not appear to be necessary for English, many English nouns, typically deverbal ones, take characteristic prepositional complements. Just as the verb "transport" is associated with the prepositions "from" and "to", the collocations "transportation from" and "transportation to" are of frequent occurrence.

= (preposition)  
FT = infinitive complement  
PI = poss-ing (gerundive) complement  
TH = "that"-clause complement  
WH = subordinate question complement

FC <form of complement> indicates the kind of syntactic construction which can function as the complement of a noun.

PP = prepositional phrase  
CP = complement phrase or clause

PC <homograph> marks nouns which have the same spelling as another part of speech. The values indicate the syntactic class of the form matched.

ADV = adverb ("a daily", "an out", "an aside")  
ADJ = adjective ("some Chinese", "a brave", "the dark")  
INF = infinitive ("the dance", "a race", "some water")  
PRPL = gerund (e.g., the noun "the painting <which we bought>" is homographic with the gerund "the painting <of the living room walls>"; "the building <on the corner>", with "the building <of the house>")  
PAPL = past participle of a verb  
OTR = other (e.g., the possessive pronoun "mine", the modal auxiliaries "will" or "might")

Although this feature is not used by the system at present, it has been included in the current lexical entries as the basis for disambiguation rules to be incorporated into the finished system.

DF <derivable from> is used with derived nominals, usually deverbal forms such as gerunds and agentive nouns, but also sometimes with nouns derived from adjective stems or other nouns. This feature particularizes the word underlying the nominal in case the noun may have 'inherited' some of the grammatical characteristics of that word. (A rough-and-ready rule of thumb is that DF <derivable from> is applicable only when the noun being coded has more letters or sounds or syllables than the related word.)

AD-A097 896

TEXAS UNIV AT AUSTIN LINGUISTICS RESEARCH CENTER

F/G 5/7

THE METAL SYSTEM. VOLUME I AND VOLUME II. APPENDICES. (U)

JAN 81 W P LEHMANN, W S BENNETT, J SLOCUM

F30602-78-C-0132

UNCLASSIFIED

RADC-TR-80-374-VOL-1/2

NL

3 in 3  
A1  
202-76 94

END  
DATE  
FILED  
5-81  
DTIC

VI = intransitive verb ("existence" from "exist")

VT = transitive verb ("government" from "govern", "gift" from "give")

VR = reflexive verb ("perjury" from "perjure")

ADJ= adjective ("happiness" from "happy", "naivete" from "naive")

NO = noun ("sisterhood" from "sister", "princeling", and "princess" from "prince")

This feature will later be used to evaluate the feasibility of incorporating derivational morphology into the grammars to a greater extent than at present. Such forms would thus not be included in the lexicon but would be derived from their base underlying forms.

ON <onset> or quality of initial sound, is required for selection of "a" or "an". Except for words beginning with h- or u-, the value is highly predictable from the spelling and so can be automated rather than hand coded for most lexical entries.

CO = consonantal

VO = vocalic

## GERMAN PREPOSITION FEATURES

CAT = PREP

*CAN	canonical form
*ALO	allomorph
RO	'case' role of prepositional phrase
NU	grammatical number of object
GC	governs grammatical case
PO	position
CN	contraction
ON	onset quality
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

CAN <canonical form> specifies the lexico-semantic element in question. The value is the basic uncontracted, unbound form of the preposition. E.g., "von" and "vo" (as in the contracted form "vom") both have "von" as the CAN to indicate that these are forms of the same preposition.

ALO <allomorph> has as a value the lexical string itself. Although the CAN <canonical form> and ALO are usually the same for most prepositions, there are several which have variable forms. For example, the "von" and "vo" mentioned above. Whenever CAN <canonical form> and ALO have the same value, ALO may be omitted in coding, as it can be supplied automatically. E.g.:

"in" CAN (IN)  
ALO (IN) - same as CAN

"i" CAN (IN)  
ALO (I) - distinct from CAN  
CN (M) - see CN <contracted> below

RO <role of argument> indicates the semantic function of the argument formed by the prepositional phrase. This feature serves disambiguation purposes and provides a consistent means of indicating preposition senses across languages. Thus the preposition "in" could be used to indicate either location within a fixed area (PLACE) or motion from one place to another (DESTINATION). Since there is not always complete overlap of functions in translating prepositions, this feature permits the selection of those translations which have the best overlap with the probable sense of the source language prepositions.

NU <grammatical number> has been introduced as a means for indicating those prepositions which restrict the grammatical number of the argument which follows. E.g., "zwischen" must have a plural object. The feature prevents analysis of "zwischen Y" as a prepositional phrase when it is part of the construction "zwischen Y und Z".

SG = singular object only  
PL = plural object only  
SG PL = no restriction [DEFAULT]

GC <governed case> specifies which case or cases are required for the object of the preposition.

G = genitive  
D = dative  
A = accusative

PO <position> marks the location of the preposition with respect to the governed nominal.

PRE = pre-posed  
POST = post-posed

CN <contraction> is used as a subscript with those prepositions which occur in contracted form. The value is that of the consonant which may follow. For example, the preposition "zu" has two values for CN: R and M.

M = M follows  
R = R follows  
S = S follows

ON <onset> identifies the initial element of the preposition. It is employed as a means for indicating the allomorph which occurs in "da" derivative forms since "r" is introduced before a preposition beginning with a vowel (e.g., "darum").

C0 = consonantal  
V0 = vocalic

## ENGLISH PREPOSITION FEATURES

CAT = PREP

*CAN	canonical form
*ALO	allomorph
RO	'case' role of prepositional phrase
NU	grammatical number of object
PO	position
ON	onset quality
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

RO <role of prepositional phrase> is used to indicate the semantic function of the prepositional phrase. Thus each sense of a preposition would have a different role function for the following noun phrase. (A complete list of role functions may be found in Appendix A-3.)

NU <grammatical number> is used primarily to distinguish those prepositions which place restrictions on the grammatical number of the following noun phrase. Thus "between" and "among" would only accept a plural or mass noun phrase, whereas most other prepositions may take either singular or plural.

SG = singular  
PL = plural  
MA PL = either mass or singular  
SG PL = either singular or plural [DEFAULT]

PO <position> marks the location of the preposition with respect to the governed nominal.

PRE = pre-posed to the NP or ADV  
POST = post-posed to the NP or ADV

ON <onset> or quality of initial sound, is required for selection of "a" or "an". Although determiners do not typically precede a preposition, they can occur in this environment when a quotation, title, or semi-idiomatic expression follows.

C0 = consonantal  
V0 = vocalic

## GERMAN PRONOUN FEATURES

CAT = PRN

*CAN	canonical form
*ALO	allomorph
CA	grammatical case
NU	grammatical number
GD	grammatical gender
PS	grammatical person
PO	position
KP	kind of pronoun
TY	semantic type
SX	sex (natural gender)
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

CAN <canonical form> is the nominative form of the pronoun, e.g., "ich", "wir", "er", "man", "was".

ALO <allomorph> is entered if the form being coded differs from the canonical form of the word. For example, the pronoun "ich" would require four separate entries with the following values for CAN <canonical form> and ALO:

CAN	ich	ich	ich	ich
ALO	ich	meiner	mir	mich

CA <grammatical case>

N	=	nominative
G	=	genitive
D	=	dative
A	=	accusative

NU <grammatical number>

SG = singular  
PL = plural

GD <grammatical gender>

M = masculine  
F = feminine  
N = neuter

PS <person> need only be coded for personal pronouns of the first and second person, the third person being a default value.

1 = first person  
2 = second person  
3 = third person [DEFAULT]

PO <position> is required for possessives of pronouns since there may be two distinct possessive allomorphs, each restricted to a specific syntactic location. The values indicate the position which the allomorph may take.

BEFMFC = before modificand  
AFTCOP = after copula

KP <kind of pronoun> lists the one or more subcategories to which the pronoun in question may belong. This in turn indicates the potential functions of the pronoun form and determines which rules may subsequently be applied.

IND = indefinite pronoun  
INT = interrogative pronoun  
REL = relative pronoun  
PER = personal pronoun  
REF = reflexive pronoun  
REC = reciprocal pronoun  
{DEM} = demonstrative  
{POS} = possessive

Note that DEM and POS are not assigned in coding PRN pronouns, rather, they are introduced in the grammar from DET determiners which function as pronouns. They originate as values for KD <kind of determiner>.

TY <semantic type> represents the semantic class of the referent of the pronoun. It is primarily used to distinguish those pronouns such as "ich", "jemand", "man", which refer to human beings, from those which reference non-sapients (i.e., the "who/which" distinction in English). The values are those of the semantic type matrix in Appendix A-2.

SX <sex - natural gender> is used for those pronouns which may refer to nouns with inherent sexuality.

M = male  
F = female  
N = neuter [DEFAULT]

## ENGLISH PRONOUN FEATURES

CAT = PRN

*CAN	canonical form
*ALO	allomorph
CA	grammatical case correlation
NU	grammatical number
PS	grammatical person
PO	position
KP	kind of pronoun
TY	semantic type
SX	sex (natural gender)
ON	onset
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

CAN <canonical form> is the form which would be used as a nominative singular subject. All paradigmatic forms of a pronoun have the same CAN value.

ALO <allomorph> is entered if the form being coded differs from the canonical form of the word. For example, the pronoun "I" would require four separate entries with the following values for CAN <canonical form> and ALO:

CAN	I	I	I	I
ALO	I	mine	my	me

CA <grammatical case correlation> is analogous to case in German in that it indicates the syntactic function of the pronoun.

S = subjective (occurs in subject position)  
O = objective (occurs as an object)  
P = possessive

NU <grammatical number> is used to indicate number distinctions.

SG = singular  
PL = plural

PS <grammatical person>

1 = first person  
2 = second person  
3 = third person

PO <position> is required for possessives of pronouns since there may be two distinct possessive allomorphs such as "her/hers", each taking a different syntactic location. The values indicate the position which the allomorph may take.

BEFMFC = before modificand  
AFTCOP = after copula

KP <kind of pronoun> is an indication of the syntactic use of the pronoun.

IND = indefinite pronoun  
INT = interrogative pronoun  
REL = relative pronoun  
PER = personal pronoun  
REF = reflexive pronoun  
REC = reciprocal  
DEM = demonstrative  
POS = possessive

TY <semantic type> characterizes the referent(s) possible. The values are those found in the semantic type matrix for nouns. (See Appendix A-2.)

SX <sex - natural gender> is used for those pronouns which may refer to nouns with inherent sexuality.

M = male  
F = female  
N = neuter (unmarked)

ON <onset> or quality of initial sound, is required for selection of "a" or "an". Although determiners do not typically precede a pronoun, they can occur in this environment when a quotation, title, or semi-idiomatic expression follows.

CO = consonantal  
VO = vocalic

## GERMAN VERB FEATURES

CAT = VST

*CAN	canonical form
*ALO	allomorph
CL	inflectional class
{PS}	grammatical person
{NU}	grammatical number
{TN}	tense
{MD}	mood
{PF}	predicate (paradigmatic) form
PX	prefix
RA	'case' role of argument
TA	semantic type of argument
MA	surface marker of argument
FA	syntactic form of argument
TT	transitivity type
VC	voice
FR	syntactic frame
AX	auxiliary
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix. Those set off by braces are introduced by the morphological and syntactic rules, largely on the basis of CL <inflectional class> values.)

DISCUSSION: The most complex of the syntactic classes which must be considered in any MT system are the verbs. They must agree with their subjects with respect to number and semantic class, and with their objects with respect to case, number, and semantic class. Verbs also carry markers for tense, mood, aspect, and voice. But a predicate may consist of more than one verbal element, non-finite verb forms such as infinitives and gerunds as well as finite verb forms (those which carry tense and number markers). In the LRC MT system the distinction is made between verbs which may function as either finite or non-finite forms,

and those which may only be finite. The latter category, consisting of those verbs most of which have traditionally been considered modals, will be subsumed under the category MDX.

CAN <canonical form> is the 'main entry' for the verb. The infinitive form of the verb is the value. If a verb lacks an infinitive, the third person singular present would be used instead.

AL0 <allomorph> is the actual surface string analyzed by the entry.

CL <morphological class> is used to specify the inflectional affixes which may occur with a particular verb stem. The values for CL are a list of numbers and letters, which identify the correct pattern of endings for the AL0 <allomorph> in each of the following categories:

present indicative	
past indicative	
present subjunctive	inflected
past subjunctive	
imperative	
- - - - -	
past participle	uninflected
infinitive	

Forms which inflect for person, number, tense, and mood have been assigned numeric values, while uninflected forms have been given alphabetic values. A complete chart of the verb classes and their affix patterns is listed below.

In coding, each AL0 <allomorph> should be assigned at least one and not more than seven values for CL, depending upon how many categories are relevant to the particular stem form being coded.

For example, the verb "lieben" is weak and has only one stem form, "lieb", which would receive the following values for CL:

AL0	CL
lieb	(PRI-1 PAI-1 PAI-1 PAS-1 IMP-1 PP-get INF-en)

The strong verb "kommen", however, has three stem forms, "komm", "kam", and "kaem", each coded separately as follows:

AL0	CL
komm	(PRI-1 PRS-1 IMP-1 PP-geen INF-en)
kam	(PAI-3)
kaem	(PAS-3)

The verb "senden" has strong and weak forms which overlap:

ALO	CL
send	(PRI-3 PAI-2 PRS-1 PAS-2 IMP-2 PP-geet INF-en)
sand	( PAI-1 PP-get )

In general, highly regular weak verbs will have one ALO <allomorph> with seven CL values, while strong and irregular verbs will have several ALO's, some of which may take only one or two values.

#### Present Indicative

CLASS	1s	2s	3s	1p	2p	3p	example verb stems
PRI-1	e	st	t	en	t	en	lieb, komm
PRI-2	e	st	t	n	t	n	handel, wander
PRI-3	e	est	et	en	et	en	antwort, bind
PRI-4	e	est/t	t	en	t	en	aechz
PRI-5	e	t	t	en	t	en	heiss
PRI-6	e	~	~	en	t	en	ess
PRI-7	e	~	~	en	et	en	fecht
PRI-8	e	~	~	~	~	~	handl
PRI-9	Ø	st	Ø	~	~	~	bedarf, vermag
PRI-10	Ø	t	Ø	~	~	~	weiss
PRI-11	~	st	t	~	~	~	baeck
PRI-12	~	st	Ø	~	~	~	gilt
PRI-13	~	est/st	t	~	~	~	drisch
PRI-14	~	t	t	~	~	~	iss
PRI-15	~	Ø	Ø	~	~	~	birst
PRI-16	~	~	~	en	t	en	wiss

#### Past Indicative

CLASS	1s	2s	3s	1p	2p	3p	example verb stems
PAI-1	te	test	te	ten	tet	ten	lieb, konn, erkann
PAI-2	ete	etest	ete	eten	etet	eten	antwort, acht
PAI-3	Ø	st	Ø	en	t	en	kam, sah, ging
PAI-4	Ø	st/est	Ø	en	t	en	trog
PAI-5	Ø	st/est	Ø	en	et	en	fand, vermied, glitt
PAI-6	Ø	est	Ø	en	et	en	trat, stand, erriet
PAI-7	Ø	est	Ø	en	t	en	schloss, entriss
PAI-8	Ø	t/est	Ø	en	t	en	blies

### Present Subjunctive

CLASS	1s	2s	3s	1p	2p	3p	example verb stems
PRS-1	e	est	e	en	et	en	lieb, antwort, komm
PRS-2	e	st/est	e	n	t	n	handel, wander, wechsel
PRS-3	e	est	e	~	et	~	laechl
PRS-4	e	~	e	~	~	~	handl, wandr, wechsl
PRS-5	~	~	~	n	~	n	laechel

### Past Subjunctive

CLASS	1s	2s	3s	1p	2p	3p	example verb stems
PAS-1	te	test	te	ten	tet	ten	lieb, sag, renn, haet
PAS-2	ete	etest	ete	eten	etet	eten	antwort, acht, send
PAS-3	e	est	e	en	et	en	kaem, ging, blies, goess
PAS-4	e	st/est	e	en	t/et	en	sah, schwuer

### Imperative

CLASS	2s	2p	example verb stems
IMP-1	ø/e	t	sauf, mach, komm, tu
IMP-2	ø/e	et	rat, bind, gleit
IMP-3	e	t	wiss, verwunder
IMP-4	e	et	rechn
IMP-5	e	~	handl, wandr
IMP-6	ø	~	befiehl, iss, gilt
IMP-7	~	t	ess, brech, seh
IMP-8	~	et	fecht, tret

### Past Participle

CLASS		example verb stems
PP-T	t	erkann, besag, durchsetz
PP-ET	et	beobacht
PP-EN	en	gelung, vermied, besess
PP-N	n	verta
PP-GET	ge-t	brann, konnt, sag, waehl
PP-GEET	ge-et	antwort, acht, send
PP-GEEN	ge-en	gang, gess, stand, seh
PP-GEEN	PP-GEN	schrie
PP-GEN	ge-en/n	ta

## Infinitive

CLASS		example verb stems
INF-EN	en	lieb, komm, geh, renn
INF-N	n	handel, wander, laechel

## PS <grammatical person>

1 = first person  
2 = second person  
3 = third person  
NIL = unmarked

## NU <grammatical number>

SG = singular  
PL = plural  
NIL = unmarked

## TN <tense>

PR = present  
PA = past  
NIL = unmarked

## MD <mood>

IND = indicative  
SUB = subjunctive  
IMP = imperative  
IRR = irrealis (contrary to fact)  
NIL = unmarked

PF <predicate form> refers to the constituent function of the paradigmatic form in question.

FIN = finite verb  
INF = infinitive  
PAPL = past participle  
PRPL = present participle

PX <prefix> marks the many German verbs that can occur with pre-posed affixes. It is necessary to enumerate which prefixes may be used with such verbs, and numerous values occur since each prefix has its own value. Most of the values are identical to the surface form of the prefix in question, for example "auf", "herab", "fern", "aus", etc. The value NIL is also used to indicate a null value or optionality.

RA <role of argument> represents the semantic functional relationship between the argument(s) present and the verb. (A list of potential values and additional description of this feature may be found in Appendix A-3.)

TA <type of argument> represents the semantic class to which the argument(s) associated with the verb must belong. For example, a verb such as "denken" must have a sentient agent. The values are thus the same as those associated with TY (semantic type) in nouns (see Appendix A-2).

MA <surface marker of argument> denotes the grammatical case, preposition, or syntactic construction which can be used to indicate the particular arguments.

N	= nominative
G	= genitive
D	= dative
A	= accusative
TH	= "dass" complement construction
FT	= infinitive complement construction
WH	= subordinate question as complement
als	= "als" phrase
	= (any prepositions, the value being their canonical form)

The value NIL may be used to indicate optionality.

FA <constituent form of argument> is used to stipulate the syntactic form of any constituents which function as complements of the verb. A wide variety of values may occur. Some verbs such as "glaub" may take a clause as object. Others require a noun phrase in a particular case, generally accusative or dative. In addition, German has verbs which take prepositional objects, for example "gelt". The values usually encountered include:

NP = noun phrase  
 PP = prepositional phrase  
 CP = complement phrase or clause  
 ADJ = adjective  
 ADV = adverb  
 PRN = pronoun  
 NIL = optional (no constituent)

TT <transitivity type> describes properties associated with voice and the combinations of arguments which typically co-occur with the particular verb form. Although there may appear to be some functional overlap here with information included by the syntactic frame values, one should note that FR <syntactic frame> determines subject and objects, while TT specifies which of several frames may be applied. In addition, it is used to determine the relationships between deverbalive nominals or adjectives and any associated complements or modificands. The most common values include:

T2AT = takes an obligatory object which is not a reflexive pronoun  
 I2AX = takes an object which must be a reflexive pronoun  
 T2AT I2AX = takes an obligatory object which may or may not be reflexive.  
 T3ATR = takes both a direct and an indirect object which is a recipient  
 T3ATB = takes both a direct and an indirect object which is a benefactive  
 T3ATX = takes at least two objects, one of which must be reflexive and one which is not reflexive  
 I1T = intransitive form with target as subject  
 I1A = intransitive form with agent as subject  
 C2AQ = copula type verbs which take an object but do not passivize  
 C2AJ = takes an adjectival complement

VC <voice> is determined by the syntactic rules on the basis of the various verb forms which are present in the predicate. A verb is assigned the value P only if it forms predicates consisting of "werden" plus a past participle for the passive.

A = active  
P = passive

FK <syntactic frame> indicates whether the entire sentence might be considered active vs. passive. Active frame sentences have an agent or instrument as subject, whereas passive frame sentences take a target, benefactive, or recipient as subject. The values indicate the potential paths which may be taken by the frame processor in identifying subjects and objects.

A = active frame  
P = passive frame

AX <form of auxiliary> specifies whether the verb takes a form of "haben" or of "sein" as its auxiliary.

HABEN = haben  
SEIN = sein

## ENGLISH VERB FEATURES

CAT = VST

*CAN	canonical form
*ALO	allomorph
CL	inflectional class
{PS}	grammatical person
{NU}	grammatical number
{TN}	tense
{MD}	mood
{PF}	predicate (paradigmatic) form
PX	prefix (adverbial particle)
RA	'case' role of argument
TA	semantic type of argument
MA	surface marker of argument
FA	syntactic form of argument
TT	transitivity type
VC	voice
OO	object order
PM	prefix mobility
ON	onset (quality of initial sound)
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix. Those set off by {} are brought in by means of the CL <inflectional class> feature during processing rather than coded separately.)

CAN <canonical form> is the infinitive form of the verb. Should no infinitive form exist, the third person singular present may be used.

ALO <allomorph> is the actual surface realization of the string covered by the entry.

CL <morphological class of the stem> serves to identify the possible set of morphological endings which a given verb may take. (The morphological phase of processing determines person, tense, and number of the verb by comparing the affix with the appropriate affix dummy.) The class values are the actual affixes, preceded by acronyms to indicate which of the five possible morphological forms they pertain to.

Infinitive	PPresent (3rd sing.)	Gerund (pres. part.)	PAst (preterit)	Participle (past)
I-Ø	PR-S	G-ING	PA-ED	P-EN
I-Y	PR-IES	G-YING	PA-IED	P-IED
I-E	PR-ES	G-BING	PA-BED	P-BED
	PR-SES	G-DING	PA-DED	P-DED
	PR-ZES	G-GING	PA-GED	P-GED
		G-KING	PA-KED	P-KED
		G-LING	PA-LED	P-LED
		G-MING	PA-MED	P-MED
		G-NING	PA-NED	P-NED
		G-PING	PA-PED	P-PED
		G-RING	PA-RED	P-RED
		G-SING	PA-SED	P-SED
		G-TING	PA-TED	P-TED
		G-ZING	PA-ZED	P-ZED
			PA-D	P-D
			PA-E	P-N
			PA-Ø	P-Ø
				P-ED

Examples of CLASS value clustering

volv-	CL(I-E)	PR-ES	G-ING	PA-ED	P-ED)
reach, miss, focus	CL(I-Ø)	PR-ES	G-ING	PA-ED	P-ED)
soar	CL(I-Ø)	PR-S	G-ING	PA- ED	P-ED)
rub	CL(I-Ø)	PR-S	G-BING	PA-BED	P-BED)
stud	CL(I-Ø)	PR-S	G-DING	PA-DED	P-DED)
plug	CL(I-Ø)	PR-S	G-GING	PA-GED	P-GED)
panic, frolic	CL(I-Ø)	PR-S	G-KING	PA-KED	P-KED)
pal	CL(I-Ø)	PR-S	G-LING	PA-LED	P-LED)
stem	CL(I-Ø)	PR-S	G-MING	PA-MED	P-MED)
don	CL(I-Ø)	PR-S	G-NING	PA-NED	P-NED)
stop	CL(I-Ø)	PR-S	G-PING	PA-PED	P-PED)
blur	CL(I-Ø)	PR-S	G-RING	PA-RED	P-RED)
gas	CL(I-Ø)	PR-SES	G-SING	PA-SED	P-SED)
knit	CL(I-Ø)	PR-S	G-TING	PA-TED	P-TED)
quiz	CL(I-Ø)	PR-ZES	G-ZING	PA-ZED	P-ZED)
cr-, impl-	CL(I-Y)	PR-IES	G-YING	PA-IED	P-IED)
cut	CL(I-Ø)	PR-S	G-TING	PA-Ø	P-Ø)
agree	CL(I-Ø)	PR-S	G-ING	PA-D	P-D)
show	CL(I-Ø)	PR-S	G-ING	PA-ED	P-N)
read	CL(I-Ø)	PR-S	G-ING	PA-Ø	P-Ø)
see	CL(I-Ø)	PR-S	G-ING	P-N)	
eat	CL(I-Ø)	PR-S	G-ING	P-EN)	
giv-	CL(I-E)	PR-ES	G-ING	P-EN)	
begin	CL(I-Ø)	PR-S	G-NING)		
swear, unwind	CL(I-Ø)	PR-S	G-ING)		
mak-, weav-	CL(I-E)	PR-ES	G-ING)		
made, unwound	CL(PA-Ø)	P-Ø)			
wove	CL(PA-Ø)	P-N)			
swor-	CL(PA-E)	P-N)			
began, ate, gave	CL(PA-Ø)				
lain, begun	CL(P-Ø)				

PS <person>

1 = first person  
2 = second person  
3 = third person  
NIL = unmarked

NU <grammatical number of the verb>

SG = singular  
PL = plural  
NIL = unmarked

TN <tense>

PA = past  
PR = present  
NIL = unmarked tense (i.e., infinitive form)

MD <mood>

IND = indicative  
SUB = subjunctive  
IMP = imperative  
IRR = irrealis (contrary to fact)  
NIL = unmarked

PF <predicate form> is introduced in the morphological phase of the translation algorithm. It serves to indicate the particular paradigmatic form of the verb.

FIN = finite verb  
INF = infinitive  
PAPL = past participle  
PRPL = present participle

PX <prefix> is used to indicate the adverbial particles which a verb may take. (These are analogous to the separable prefixes in German.) The values are the adpreps themselves, for example:

out (as in "ask out")  
over (as in "run over")  
up (as in "look up")  
down (as in "break down")  
NIL (no prefix)

RA <role of argument> represents the functional relationship between the verb and each argument with which it is associated. The values are semantic relations and are used to determine subject selection and the sequence of the arguments in the sentence. Some typical values are given below as examples. A complete list of potential values may be found in Appendix A-3.

AGT = agent  
TAR = target  
BEN = benefactive  
REC = recipient  
INS = instrument  
LOC = locative (any locative)  
PLC = place (no motion)  
DST = destination or goal (direction to)  
ORN = origin (direction from)

TA <type of argument> the values associated with this subscript are the usual semantic categories of nouns and pronouns. Any of the potential semantic type values (see Appendix A-2) may be used to describe an argument a verb requires.

MA <syntactic marker of argument> is used to indicate syntactic markers which may be used to identify the role of an argument in the sentence. The values are usually prepositions or U (English unmarked case), although pronouns may be marked for object and subject relations.

— = (any prepositions, the value being their canonical form)  
S = subjective  
O = objective  
U = unmarked  
TH = "that" complement  
FT = "for-to" (infinitive) complement  
PI = "poss-ing" (gerundive) complement  
WH = complement initiated by a "WH" word (who, what, when, etc.)

FA <constituent form of argument> indicates the syntactic form of the constituent which must comprise the argument. The values may be any non-terminal node derived by the system. Usual values might include the following:

NP = noun phrase (used as subject or object)  
PP = prepositional phrase (where a preposition given under MA indicates the function of the argument)  
CP = complement phrase or clause  
ADV = adverb (used if the verb takes an obligatory adverbial argument)  
PRN = pronoun (typically used when "it" is the indefinite subject)  
ADJ = adjective (used for adjectival complements, as with sensory verbs, e.g. "smell good")  
NIL = optionality

TT <transitivity type> describes properties associated with voice and the combinations of arguments which typically co-occur with the particular verb form. It also predicts passivization and the semantic properties of complements and modificands which may be associated with deverbal nominal or adjectives. The more common values include:

T2AT = takes an obligatory object which is not a reflexive pronoun  
I2AX = takes an object which must be a reflexive pronoun  
T2AT I2AX = takes an obligatory object which may or may not be reflexive  
T3ATR = takes both a direct and an indirect object, which is a recipient (marked by "to")  
T3ATB = takes both a direct and an indirect object which is a benefactive (marked by "for")  
T3ATX = takes at least two objects, one of which must be reflexive and one which is not reflexive  
ILT = intransitive form with target as subject

I1A = intransitive form with agent as subject

I2AL = intransitive verb which takes a locative complement (e.g., come, go, etc.)

C2AQ = copula type verb which takes an object but does not passivize (e.g., "weigh", "cost", etc.)

C2AJ = takes an adjectival complement (e.g., "smell", "seem", etc.)

VC <voice> is determined by the syntactic rules on the basis of the various verb forms which are present in the predicate. The values given in the lexical entry are those which may potentially be assigned to the verb. The values determine how the verb is processed by the frame processor.

A = active frame  
 P = passive frame

OO <object order> refers to the order of the direct object and the indirect object in relation to each other. Most verbs permit either to precede the other:

He bought the book for Mary.  
 He bought Mary the book.

Others are restricted:

He purchased the book for Mary.  
 \*He purchased Mary the book.

The values are:

DI = direct-indirect only  
 ID = indirect-direct only  
 DI ID = either order [DEFAULT]

PM <prefix mobility> is used to indicate the positions which may be taken by the adverbial particle analogous to a German prefix.

BEFOBJ = before an object noun  
 AFTOBJ = after an object noun

Note that forms which can occur after a noun phrase object must occur before a pronominal object. This is taken into account in the grammar rules.

ON <onset> or quality of initial sound, is required for selection of "a" or "an". Although verbs are not typically preceded by determiners, deverbal adjectives (e.g., gerunds and participles) may quite readily occur in a noun phrase, and so the stems must be marked for onset. Except for words beginning with h- or u-, the value is highly predictable from the spelling and so can be automated rather than hand coded for most lexical entries.

CO = consonantal  
VO = vocalic

## GERMAN MODAL-AUXILIARY FEATURES

CAT = MDX, HABEN, SEIN, WERDEN, LASSEN

*CAN	canonical form
*ALO	allomorph
PS	grammatical person
NU	grammatical number
TN	tense
MD	mood
PF	predicate (paradigmatic) form
RA	'case' role of argument
TA	semantic type of argument
MA	surface marker of argument
FA	syntactic form of argument
TT	transitivity type
VC	voice
FR	syntactic frame
MU	use of modal
AX	auxiliary
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag
*TAG	area of provenience tag

(Asterisks mark systemic features described generally at the beginning of this appendix)

DISCUSSION: German has a variety of auxiliaries such as "haben", "moegen", "sein", "lassen", etc. The auxiliaries traditionally called "modals" are coded as CAT (MDX) while "haben" is CAT (HABEN), "sein" is CAT (SEIN), "werden" is CAT (WERDEN), and "lassen" is CAT (LASSEN). Other auxiliaries which are syntactically similar to modals may be either put in CAT (MDX) or given their own CAT, depending on their patterns. Many of the features are the same as those found with the non-finite verbs. However, additional information is specified for the modals since they appear in the lexicon in their inflected forms, whereas most other verbs are entered as stem forms and derive tense, mood, and number information from their suffixes.

CAN <canonical form> is the infinitive form of the modal, assuming one exists. If there is no infinitive, the third person singular present is employed.

PS <grammatical person>

1 = first person  
2 = second person  
3 = third person

NU <grammatical number>

SG = singular  
PL = plural

TN <tense> indicates the tense of the verb form. Since German has two sets of inflectional suffixes for tense, the two possible values are:

PR = present  
PA = past

MD <mood> for German has the associated values:

IND = indicative  
SUB = subjunctive  
IMP = imperative  
IRR = irrealis (contrary to fact)

PF <predicate form> indicates the paradigmatic function class of the form in question.

FIN = finite form  
INF = infinitive  
PAPL = past participle  
PRPL = present participle

RA <role of argument> represents the semantic functional relationship between the argument(s) present and the verb. (A list of potential values and additional description of this feature may be found in Appendix A-3.)

TA <type of argument> represents the semantic class to which the argument(s) associated with the verb must belong. For example, a verb such as "lassen" must have a living agent. The values are thus the same as those associated with TY (semantic type) in nouns (see Appendix A-2).

MA <surface marker of argument> denotes the grammatical case, preposition, or syntactic construction which can be used to indicate the particular arguments.

N = nominative  
G = genitive  
D = dative  
A = accusative  
TH = "dass" complement construction  
FT = infinitive complement construction  
WH = subordinate question as complement  
als = "als" phrase  
- = (any prepositions, the value being their canonical form)

The value NIL may be used to indicate optionality.

FA <constituent form of argument> is used to stipulate the syntactic form of any constituents which function as complements of the verb. A wide variety of values may occur. The values usually encountered include:

NP = noun phrase  
PP = prepositional phrase  
CP = complement phrase or clause  
ADJ = adjective  
ADV = adverb  
PRN = pronoun  
NIL = optional (no constituent)

TT <transitivity type> indicates the combinations of arguments which can function as subject and object/complements of the modal. The values are similar to those used for other verbs. Common values used with modals are:

T2AT = takes an obligatory direct object (use as main verb)  
I1A = intransitive form with agent as subject (use as main verb)

M2AC = takes an agent as subject and another verb form as complement (use as modal)

C2AQ = copula type verb which takes two arguments but does not passivize. The second argument may be an agent, a locative, a temporal, or predicate adjective.

VC <voice> indicates potentiality for the verb to occur as the main verb of active or passive sentences.

A = active  
P = passive

FK <syntactic frame> indicates whether the word is processed by the active frame (agent or instrument as subject) or passive frame (target as subject).

A = active  
P = passive

MU <use of modal> indicates the possible functions of the modal within a sentence.

MDX = modal auxiliary  
MVB = main verb  
COP = copula  
AUX = auxiliary  
XBE = existential "be"

AX <form of auxiliary> expresses the class of finite verbs which may be used with a non-finite form of the modal.

HABEN = haben  
SEIN = sein  
WERDEN = werden  
MDX = a modal other than "haben", "sein", or "werden"

## ENGLISH MODAL/AUXILIARY FEATURES

CAT = MDX, BE, HAVE, WILL

*CAN	canonical form
*ALO	allomorph
PS	grammatical person
NU	grammatical number
TN	tense
MD	mood
PF	predicate (paradigmatic) form
RA	'case' role of argument
TA	semantic type of argument
MA	surface marker of argument
FA	syntactic form of argument
TT	transitivity type
VC	voice
MU	use of modal
ON	onset quality
*CAT	lexical category
*PLC	placement
*PRF	preference
*LEX	lexical collocation
*SNS	sense number
*CNO	concept number
*TAG	area of provenience tag

(Asterisks mark systemic features described  
generally at the beginning of this appendix)

DISCUSSION: This is actually a group of related categories for verbs which are used as auxiliaries in English. That is, they may be followed by a non-finite form of another verb, forming a complex predicate. (For convenience, we refer to all of them as "MDX's" even though some have their own CAT <lexical category> labels.) Unlike the verbs with the category symbol VST, the MDX's are included in their fully inflected forms in the lexicon.

CAN <canonical form> is the third person singular form of the modal/auxiliary in question.

AL0 <allomorph> is the actual string represented by the entry. The various paradigmatic forms of a modal/auxiliary will have differing AL0 values but the same CAN <canonical form> value.

PS <person> specifies the distinctions of grammatical person.

1 = first person  
2 = second person  
3 = third person

NU (grammatical number)

SG = singular  
PL = plural

TN <tense> indicates the tense of the verb form. Since English has two inflectional sets of suffixes for tense, there are two possible values (in addition to the null value).

PR = present  
PA = past

MD <mood>

IND = indicative  
SUB = subjunctive  
IMP = imperative  
IRR = irrealis (contrary to fact)

PF <predicate form> indicates the paradigmatic form in question, and limits potential syntactic usage of the form.

FIN = finite form  
INF = infinitive  
PAPL = past participle  
PRPL = present participle

RA <'case' role of argument> represents the semantic function(s) of any argument(s) required by the modal/auxiliary. (Consult Appendix A-3 for a list of potential values.)

TA <semantic type of argument> designates the semantic class of potential arguments. The values are taken from the noun semantic type matrix. (See Appendix A-2.)

MA <surface marker of argument> indicates potential means for id. titiving the arguments which co-occur with the modal/auxiliary.

S = subject function (for pronouns)  
O = object function (for pronouns)  
U = unmarked (for NP's, ADV's, etc.)  
— = (any prepositions, the value being their canonical form)  
TH = "that" complement or clause  
FT = "for-to" (infinitive) complement  
PI = "poss-ing" (gerundive) complement

FA <constituent form of argument> specifies the node in the grammar which is built by the rules using the modal/auxiliary.

NP = noun phrase  
PP = prepositional phrase  
CP = complement or clause  
ADJ = adjective (used for adjectival complements)

TT <transitivity type> describes the combinations of arguments which co-occur with the modal/auxiliary.

T2AT = takes an obligatory direct object (use as main verb)  
I1A = intransitive form with agent as subject (use as main verb)  
M2AC = takes an agent as subject and another verb form as complement (use as modal/auxiliary)  
C2AQ = copula type verb which takes two arguments but does not passivize. The second argument may be an agent, a locative, a temporal, or predicate adjective.

VC <voice> specifies whether the modal/auxiliary serves as the main verb in active or in passive sentences. Most modals/auxiliaries do not passivize and so only have "A" (active) as their value. However, "have" as a main verb is an exception.

A = active  
P = passive

MU <modal/auxiliary usage> indicates potential functions of the modal/auxiliary within the sentence string.

MDX = modal auxiliary  
MVB = main verb  
COP = copula  
AUX = auxiliary  
XBE = existential "be"

ON <onset> or quality of initial sound, is required for selection of "a" or "an". Although determiners do not typically precede a modal or auxiliary, they can occur in this environment when a quotation, title, or semi-idiomatic expression follows.

CO = consonantal  
VO = vocalic

## APPENDIX A-2

### Semantic Type

Semantic co-occurrence restrictions are helpful in disambiguation of potentially ambiguous analyses. In restricted domains it is possible to achieve quality translation without resorting to semantic restrictions, but in a wider area of discourse one must further subdivide the lexicon into semantic categories. Semantic information can be inserted into the lexical entry by means of the semantic type features. Such features as TY <semantic type of noun or of referent of pronoun>, TM <semantic type of adjective modificant>, TC <semantic type of complement of noun or adjective>, TA <semantic type of argument of verb or modal/auxiliary> are introduced at the lexical level as a matrix of values. Thus each noun is individually scored for a variety of characteristics which must correlate with those of the other parts of speech it may co-occur with.

#### Physical Structure

MAT	material
ENT	entity
TAN	perceptible to senses

#### Group Relations

VAR	variable
HOM	homogenous
PRT	partial
SET	set/collective

#### Abstractions

REL	relational
ATR	attributive
QUA	quality/domain

#### Aspectuals

STA	state
ACT	activity
EVT	event

#### Life Processes

LIV	living
VOL	volitional
SEN	sentient
RSP	responsible
CPX	complex

#### Adverbial Localization

GEO	geographic/map structures
SPA	spatial approximations
TMP	temporal

#### Special Characteristics

COM	commodity
MSR	measure
DAT	data/informational

Unlike values for the other features in the system, in addition to the relevance of the attribute represented by the letters of the acronym, a preceding plus and/or minus is required to signify whether it is pertinent in a positive or negative manner:

- + = morpheme has (or requires) this attribute
- = morpheme does not have this attribute (or requires a corresponding minus)
- +/- = morpheme may or may not have (or require) this attribute, i.e., will succeed in any intersection. This differs from an attribute being irrelevant, as temporality (TMP) would be when no consideration of time is involved. When the value is irrelevant, nothing is coded, and the value is 'NIL'.

The semantic system will use intersections of plusses and minuses (NIL values will fail in any intersection). Thus a verb such as "murder" will have the value +RSP and will require (i.e., will only intersect with) a subject noun which is also +RSP. This means that nouns which are -RSP, e.g., "baby", and nouns which are not coded for RSP (=NIL), e.g., "floor", will fail as subjects for "murder". A noun coded as +/-RSP, e.g., "child", will, of course, also succeed in the intersection. Minus values (as opposed to NIL values) are useful for those instances in which lack of an attribute is important. Minus values permit a verb such as "conceive" (in the sense of 'imagine') to be coded as requiring a -MAT object noun. In this instance, any noun marked -MAT or +/-MAT will succeed; a noun marked +MAT will fail. (NIL values for MAT are not allowed, since MAT is one of the values which is obligatorily coded. If MAT is not coded, a failure for the intersection would result.)

Thus the semantic type feature (TY) for a noun such as "book", in the sense of a series of printed sheets bound together, would take the form:

TY (+MAT +ENT +TAN -VAR -HOM -SET -PRT)

The following features represent a bare minimum which must be considered to handle semantic type agreement.

PHYSICAL STRUCTURE - refers to the physical (material) composition of the noun's referent. [Obligatory for all nouns.]

Material (MAT) - These nouns represent physical objects, i.e., things which are composed of matter. Such nouns include the class which has traditionally been considered 'concrete'. Abstract nouns are 'minus material'.

Examples: +MAT = box, atom, particle, dirt, telephone, gas  
-MAT = idea, vision, concept, relationship, math, aroma

Entity (ENT) - These nouns represent distinct and separate units with identifiable boundaries. Such nouns include the classes which have traditionally been considered 'count' nouns. Some abstract nouns are 'plus entity'.

Examples: +ENT = bird, tree, lake, star, tool, idea, design  
-ENT = existence, physics, energy, importance, hydrogen  
+/-ENT = milk, bone, metal

Tangible (TAN) - These nouns represent forms which can be perceived with the traditional senses.

Examples: +TAN = bird, fragrance, flavor, ocean, heat  
-TAN = idea, existence, correlation

Example sets:

book bone oxygen dozen electron ghost aroma idea physics

MAT	+	+	+	+/-	+	-	-	-	-
ENT	+	+/-	-	+	+	+	+/-	+	-
TAN	+	+	+/-	+	+/-	+	+	-	-

---

GROUP RELATIONS - describe set characteristics of the object referred to and related nouns. [At present all of these appear to be potentially 'plus entity' and/or 'plus material', that is, they are either non-abstract or may refer to non-abstract nouns. Only the plus value is relevant; minus is equivalent to nil.]

Variable (VAR) - Such nouns represent entities which are capable of change in physical conformation. The class includes forms which are flexible or fluid.

Examples: +VAR = liquid, gas, population  
-VAR = box, mountain, tree, metal

Homogeneous (HOM) - These nouns have a uniform material composition and structure with no subdivision into distinct units. Traditional 'mass' nouns generally fall into this class.

Examples: +HOM = milk, metal, water, soil, blood, ice  
-HOM = dish, hammer, icecube, erythrocyte

Partial (PRT) - These nouns represent forms which are usually associated with a larger whole and which typically do not function independently.

Examples: +PRT = head, tip, piece, core, dial, hand

Set/Collective (SET) - These nouns act as a cover term for a more or less homogeneous group of objects. The class includes most of those nouns traditionally considered collectives, but not typically mass nouns.

Examples: +SET = group, class, collection, bunch, herd, pack

Example sets:

	book	piece	metal	milk	population	group	chlorine
VAR	-	-	-	+	+	+/-	+
HOM	-	-	+	+	-	-	+
SET	-	-	-	-	+/-	+	-
PRT	-	+	-	-	-	-	-

ABSTRACTIONS - represent abstract attributive characteristics. [All are 'minus material'.]

Relational (REL) - These nouns represent the relationship between two or more objects. They typically take complements with "between" or "among". They are often deverbal or deadjectival forms whose corresponding verbs take conjoined subjects.

Examples: +REL = similarity, difference, correlation, marriage

Attributive (ATR) - These nouns describe an attribute or quality of a particular object or concept from the viewpoint of the observer. Most are deadjectival in form and sense.

Examples: +ATR = smallness, importance, significance, strength

Quality/Domain (QUA) - Qualities that serve to define a set of objects or concepts. The exact boundaries tend to be based upon personal opinion and are therefore subject to debate.

Examples: +QUA = art, beauty, truth, physics, mechanics

Example sets:

format similarity significance science

REL	-	+	-	-
ATR	-	+/-	+	-
QUA	-	-	-	+

---

ASPECTUALS - represent verbal or activity descriptions. [All are 'minus material'.]

State (STA) - These nouns represent a characteristic, attribute, or state of being. They are similar semantically to adjectives.

Examples: +STA = art, death, health

Activity (ACT) - These nouns include both on-going processes and single events. The class includes many deverbal nouns.

Examples: +ACT = birth, death, singing, dancing

Event (EVT) - These nouns are 'happenings'. These are similar to activities but emphasize the event rather than the process.

Examples: +EVT = concert, birth, death, wedding, graduation, game

Example sets:

	life	birth	death	dancing	concert	man
STA	+	-	+	-	-	-
ACT	-	+	+	+	-	-
EVT	-	+	+	-	+	-

Note: "man" would not need to be scored for these characteristics since it is +MAT. The minus values here are redundant.

LIFE PROCESSES - Characteristics associated with 'life'. [All = +MAT]

Living (LIV) - Forms which are traditionally considered to be alive (as opposed to dead) or animate (as opposed to static).

Examples: +LIV = man, cat, bacterium, tree, mushroom  
Verbs: Subject +LIV = live, die

Volitional (VOL) - The nouns represent forms which are spoken of as capable of independent thought and/or action.

Examples: +VOL = man, dog, alien, computer, robot  
Verbs: Subject +VOL = think, decide, plan

Sentient (SEN) - Used to characterize nouns which represent forms which possess the usual senses.

Examples: +SEN = man, dog, cat, teacher, child  
Verbs: Subject +SEN = feel, see, taste, smell, hear

Responsible (RSP) - These nouns represent entities which may be held responsible for their actions. Usually these nouns refer to adult humans, either individually or collectively.

Examples: +RSP = man, doctor, sister, police, court, jury  
Verbs: Subject +RSP = murder, sentence, insist

Complex (CPX) - These nouns represent complex entities comprised by individual identifiable parts which move in an organized manner to perform some specific act.

Examples: +CPX = man, computer, dog, automobile  
Verbs: Subject +CPX = run, die

Example sets:

man computer dog fly bacterium tree car rock

LIV	+	-	+	+	+	+	-	-
VOL	+	+	+	-	-	-	-	-
SEN	+	-	+	+	+	-	-	-
RSP	+	-	-	-	-	-	-	-
CPX	+	+	+	+	+/-	-	+	-

---

ADVERBIAL LOCALIZATION - these nouns serve to pinpoint activities and other objects in relation to time and space. In this respect they are somewhat similar to adverbs. [Only coded when relevant; minus and NIL are equivalent.]

Geographic/Map Structures (GEO) - these nouns represent forms which usually do not move from their expected location and may be used to identify the location of other objects or activities.

Examples: +GEO = mountain, city, Lake Erie, Guadalupe Street

Spatial Approximations (SPA) - These nouns represent place, but are not necessarily fixed in a particular location, rather they may vary with movement of the observer's perspective or point of reference.

Examples: +SPA = sky, north, horizon, depth

Temporal (TMP) - These nouns are typically associated with measurement of time.

Examples: +TMP = week, moment, generation, future, eon

SPECIAL CHARACTERISTICS - refer to properties which may distinguish a noun as a member of a special syntactic class. This group of features may be augmented depending upon requirements within a particular genre. [These features are only coded when relevant. A nil value is equivalent to a minus.]

Commodity (COM) - These nouns are things which are transportable but which are not necessarily physical objects.

Examples: +COM = format, news, text, tox, car

Measure (MSR) - This class includes nouns which represent defined quantities of matter, time, or energy. These units are "fixed" in that each defines a set of entities which are indistinguishable on the basis of the quality in question. The units of measurement belong in the +MSR class.

Examples: +MSR = ton, week, month, dollar, pound, year

Informational/Datum (DAT) - These nouns are capable of imparting or transmitting linguistic or numeric information.

Examples: +DAT = computer, person, text, broadcast, paper, typewriter

\*\*\*\*\*  
Abstract Noun Matrix

	COM	ACT	DAT	VAR	REL	ATR
	STA	MSR	QUA	EVT	LOC	
structural: "format"	+	-	-	-	-	-
informational: "news"	+	-	-	+	-	-
definite state: "death"	-	+	-	-	+	-
judgmental state: "art"	-	+	-	-	+	-
defined state: "physics"	-	+	-	+	+	-
single event: "birth"	-	-	+	-	-	-
ongoing: "singing"	-	-	+	-	-	-
measuring: "week", "ton"	-	-	-	+	-	-
measured: "air pressure"	-	-	-	+	-	-
locational: "sky", "horizon"	-	-	-	+	-	+
relational: "similarity"	-	-	-	+	-	-
qualitative: "importance"	-	-	-	+	-	-

\*\*\*\*\*

To facilitate coding, it is possible to use brief designations which represent configurations of TY values. For a vertebrate animal other than human, for instance, one can code TY (ANI) which signifies--

+MAT +ENT +TAN -VAR -HOM -PAR -COL +LIV +/-VOL +/-SEN -RSP +CPX +COM

Or, a liquid substance may be simply coded TY (LIQ), which can later be automatically transformed into TY (+MAT +/-ENT +TAN +VAR +HOM +COM) to fit the requirements of the MT system. These quasi-values are:

ABA = abstract or animate  
ABS = abstract  
ANI = animal (not human) - sex unspecified  
ANU = animal or human (sex unspecified)  
COL = collective and animal - collective and human  
CON = concrete (not abstract) - animate or inanimate  
FAN = female animal  
FEM = female (animal or human)  
FRU = female human  
GAS = gas  
HSO = human and solid  
HUM = human - sex unspecified  
INA = inanimate (not abstract) - solid, liquid, or gas unspecified  
LIQ = liquid  
MAL = male (animal or human)  
MAN = male animal  
MHU = male and human  
MOV = movable (solid)  
NAN = not animate (abstract or inanimate)  
NCO = not concrete or animal (abstract or human)  
NHU = plant or animal (not human)  
NMO = not movable (solid)  
PHY = physical qualities not associated with actual matter-- e.g.,  
diseases, measurements or physical dimensions such as heat,  
light, or sound, as well as words like "heat" or "light"  
themselves.  
PLA = plant  
SLI = solid or liquid (not gas)  
SOL = solid (movable or not movable unspecified)  
XXX = unmarked (no semantic restrictions)

## APPENDIX A-3

### Argument Roles

The IRC translation system uses semantic 'case' role structures as a basis for describing the relationships among the various constituents of the sentence. An argument (ARG) is a noun phrase or adverb phrase which functions as a constituent within the sentence. Of the arguments directly dominated by the sentence (S) node, it is useful to distinguish two classes: those which are 'central' to the predicate and those which are 'peripheral'.

Central arguments may function as subjects or objects of a particular verb and they are specified in the lexical description of verbs. They may be marked by a preposition, though they may also be signified solely by grammatical case or by position. At least one central argument is required for every verb, since only central arguments may function as subjects.

Peripheral arguments may occur with any verb, and are not specified as obligatory constituents in most verb descriptions. They are usually introduced by prepositions which indicate the role of the resulting argument. Adverbs are also included in this category and have their role specified in their lexical descriptions.

Roles currently available in the system--

#### Major Central Roles

AGT	agent
INS	instrument
TAR	target
BEN	benefactive
REC	recipient
RFX	reflexive
TOP	topic
COM	commutative
BPT	body part

#### Major Peripheral Roles

LOC	locative (subsumes STA and MOT)
STA	station (subsumes PLC and PRX)
PLC	place
PRX	proximity

MOT	motion (subsumes ORN, DST, CIR, ART)
ORN	origin/source
DST	destination/goal
CIR	circumferential
ART	area traversed
CHR	characteristic/composition
EFF	effect/result
EQU	equitive
MAT	material
MSR	measure
MTH	method/means
PUR	purpose/reason
SIM	similarity
SOB	state-of-being
TRN	transitional

#### Additional Peripheral Roles

TMP	temporal (subsumes DUR and PUN)
PUN	punctual
DUR	durative
CTR	contrary
DSP	despite
ENV	environment/context
EXC	exception/exclusion
MAN	manner
POS	possessor
PRT	partitive
RES	respective/comparative

#### MAJOR CENTRAL ARGUMENT ROLES:

These roles are only used as central arguments and thus may only be assigned within a sentence if the verb specifies the particular argument role.

Agent (AGT) - the individual or thing initiating the action. The agent may be either animate or an inanimate force. An agentive argument is typically the subject of an active transitive verb but may also occur with some intransitives. (Other intransitives may have a target subject.) The agent can often be identified by answering the questions 'Who did it?' or 'What made it happen?'. The preposition "by" is often used to designate agency.

The kitten drank the milk.  
The barn was destroyed by the tornado.  
The poet died before publication of his work.

Instrument (INS) - the individual or thing used to perform the action. The instrument is 'used', it does not 'initiate' action. A constituent may be identified as the instrument if a paraphrase of the sentence can be written in which the term in question is the direct (accusative) object of the verb "use". When a question or relative is formed, "how" or "with what" are the usual pronouns. If the instrument is not the subject, it may typically be signified by "with", although "by" may also be used in some contexts, as may "of".

The box was opened with a crowbar.  
He used a crowbar to open the box.

Target (TAR) - The recipient of the action. The object usually undergoes a change of state as a result of the action. It may be the subject of a passive sentence.

Peter dropped the glass; the fall broke it.  
The glass was broken by the impact

Benefactive (BEN) - The individual or thing which benefits from the activity. The benefactive and purpose/reason (PUR) roles are sometimes confused in that they take similar markings, but the PUR does not function as a potential subject and cannot be used as a indirect object in a sentence position between the verb and direct object or target. When a marker is expressed, it is generally "for".

The man bought his mistress the necklace.  
The man bought the necklace for his mistress.

Recipient (REC) - The individual or thing which receives something as a result of the action. "To" often marks the recipient. It must be distinguished from destination/goal (DES), which allows other prepositions and which cannot function as subject.

The man gave his wife a bracelet.  
The man gave a bracelet to his wife.

The above central roles correspond to the traditional notions of subject, direct object, and indirect object. In a typical active transitive sentence the AGT (agent) [or INS (instrument) if AGT is not expressed] will be the subject (nominative). The TAR (target) will be

the direct (accusative) object, and the benefactive (BEN) or recipient (REC) will be the indirect (dative) object. In the passive, the TAR, BEN, or REC will become subject.

For intransitive verbs, the subject is either an AGT or a TAR. It may be necessary to examine a transitive form of the verb to distinguish these roles if it is not semantically obvious:

The cookie broke - TAR - John broke the cookie  
John ate - AGT - John ate the cookie

Reflexive (RFX) - This role has been used for certain obligatory reflexive pronouns which are not true targets. They cannot be replaced by a noun phrase or non-co-referential pronoun nor can they function as subjects.

John contented himself with second place.  
The child behaved herself today.

In the above examples, the reflexive pronoun is not a true target and is obligatorily a reflexive pronoun referring to the agent, unlike the reflexive in a sentence such as:

John saw himself in the mirror - John saw Mary in the mirror.

Reflexive pronouns used in the reflexive (non-target) role are quite rare in English, but they are common in languages such as German.

Topic (TOP) - The thing described. It is often used with "on" or "about", although many verbs do not require a preposition. The subject of "be" plus a predicate is often a topic.

We talked about the plans.  
The book is here.

Comitative (COM) - This is used to represent co-agency and co-objectivity. It is an unusual situation, but so is the group of verbs which use it. There is a problem in that, unlike languages such as Arabic which have distinct verb forms for reciprocal action, a wide variety of syntactic structures may be used for it in English and German. In many instances it can be represented by a conjoined subject, often with a phrase like "each other" in the object slot. Or the relationship can be identified by "with".

Joseph and Mr. Smith shook hands,  
John and Mary kissed each other,  
The Johnsons are sitting with the Smiths.

Body Part (BPT) - There may be some question as to whether body parts are subsets of other roles, such as instrument, or whether they represent a distinct role. Since such forms seem to behave in anomalous ways, it may be easier to analyze all body parts as unique category of their own which can be recognized on the basis of the semantic type of the head noun.

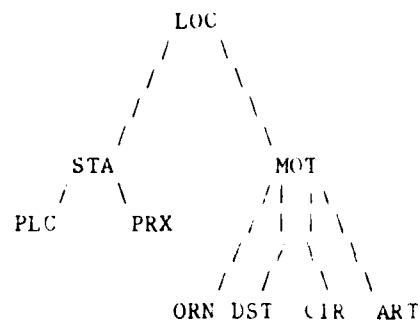
John covered his eyes with his hand.  
My head aches.

They probably do not require separate role assignation, but rather the NP's with such nouns as head word may participate in particular syntactic rules. Their sentential function should fall within the usual roles. This would be similar to the use of nouns such as "tip" and "piece".

#### PERIPHERAL ARGUMENT ROLES WHICH MAY BE INCLUDED IN VERB DESCRIPTIONS:

These roles are typically peripheral and thus may be used with any verb. However, they are also specified in the argument descriptions for certain classes of verbs, most notably the verbs of motion, which take locative ARGs. In such usage, they fall within a classification similar to central roles but they are not eligible for subject formation, hence the dual specification.

Nine locative arguments are inter-related so as to form a tree in which a match is possible between any lower node and a higher on the same branch.



Locative (LOC) - Marks the location of an action, of an object or an individual. It may indicate a fixed position as station (STA) does, or direction or course as motion (MOT) does, and hence subsumes those roles. Almost any preposition can mark a locative.

Station (STA) - Indicates the static location of an object or activity. It subsumes place (PLC) and proximity (PRX).

Place (PLC) - The locale of the action or object.

The book is on the table.

Proximity (PRX) - Indicates adjacent objects or proximity to an object or action. Typical prepositions are "by", "near", "next to", "beside".

The book is next to the telephone.  
There is a tavern nearby.

Motion/Mobile (MOT) - Specifies the direction of motion in an activity. It subsumes origin/source (ORN), destination/goal (DST), circumferential (CIR), and area traversed (ART).

origin/Source (ORN) - Place from which. "From" is the typical marker.

The man came from New Haven.

destination/goal (DST) - Place to which. "To" and "toward" are frequently used.

The girl went to the store.

Circumferential (CIR) - Used to indicate location about a (relatively) fixed point. "Around" and "about" are the most common markers.

The earth rotates on its axis.  
The moth flew around the candle.

Area Traversed (ART) - Used to indicate location transversed.  
"Through" and "between" are typical prepositions.

The girl ran through the trees.  
We traveled via New York.

her potentially verb-related peripheral arguments do not operate in tree structure.

Characteristic/Composition (CHR) - A property which distinguishes an individual or thing. Like possessor, it is identified by the genitive case, the preposition "of", or verbs such as "have".

The book of poems....

Effect or Result (EFF) - Indicates the result or consequence of some action. One of its uses is in predicate complement constructions, such as those following causative verbs.

He made her eat the cake.

Equitive (EQU) - This might be used to indicate equality or a characteristic of the subject. It would be typically used for the object of verbs such as "be", "become", and "seem". It is sometimes possible to handle this relationship under other roles such as characteristic, locative, partitive, etc. It has been suggested that the role EQU be introduced as a cover term for the various constituents which can fill this slot.

Material (MAT) - Used to express composition with verbs of creation. It may answer the question 'how?' or 'with what?' and the usual prepositions are "with", "out of", "from". Superficially MAT is somewhat similar to origin/source (ORN) in terms of marking, but occurs with a different class of verbs.

Mary made the cake from a mix.  
John built the wall out of bricks.

Measure (MSR) - Used to indicate quantities. Usually MSR role constituents can be recognized on the basis of head nominals which have measurement as semantic type, although certain verbs may also provide clues. Verbs such as "weigh" and "cost" typically take a MSR argument.

The fish weighed seven pounds.  
She sang for an hour.

Method/Means (MTH) - The means or mechanism by which an action is carried out. This is typically an abstract or non-tangible nominal. It may be represented by "by" or "with". "From" and "of" are also possible. The method may be used as subject of the phrase 'used as a means of....'

Bacteria reproduce by binary fission.

Purpose/Reason (PUR) - The motivation behind the action, either from prior causes or desired goals. The role answers the question 'why?' and typically takes "from" or "for" along with a variety of phrasal structures such as "because of", "due to" or "as a result of".

She suffered from boredom.  
He'd do anything for money.  
He resigned because of ill health.

Similarity (SIM) - Indicates identity or resemblance to another entity. These arguments fill the same syntactic and logical functions as adjectives, and the verbs which take SIM are often verbs which take adjectival complements.

She looked like her mother.

State of Being (SOB) - Indicates a state which characterizes the referent.  
He is on duty/at war/under age/at our mercy.

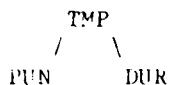
Transitional (TRN) - Used to indicate a change of state.

Heat can change water into steam.  
His mood went from grave to gay.

#### OTHER PERIPHERAL ARGUMENT ROLES:

These roles are only used peripherally, that is, they are never specified in the lexical description of a particular verb. They may occur with any verb and serve to provide additional information or they may be used as complements of noun phrases.

Three of these peripheral arguments relate to time and form a tree.



Temporal (TMP) - The time when the action takes place. Adverbs such as "yesterday" fill this slot. It answers the question 'when?' and subsumes punctual (PUN) and durative (DUR).

Punctual (PUN) - The time at which something takes place. Prepositions such as "before" and "after" mark nominals. It marks 'perfective' as opposed to 'imperfective' reference in the temporal.

We will go after the movie.

Durative (DUR) - This is the time span within which something takes place. It is used to mark 'imperfective' as opposed to the 'perfective' of the temporal. Although it is probably not usually necessary to distinguish the two, presence of one or the other may influence the verb form in some instances and so it is advisable to allow the option of distinguishing at least two time frame roles. We may not need them for English and German, but some languages do have restrictions in terms of what adverbial phrases can be used with various verb aspects.

It rained during the night.

Other peripheral arguments do not operate in a tree structure.

Contrary (CTR) - Indicates that one argument is incompatible with another.

He did it against my wishes.

Despite (DES) - Used to indicate an oppositional force, as is marked by "despite", "in spite of", "regardless of", etc.

He came despite bad weather.

Environment/Context (ENV) - The context in which an action is taking place. This is essentially a non-locative locale.

He heard her above the din.  
That is one among many.

Exception/Exclusion (EXC) - Indicates an exception to or exclusion from the action of the verb, an argument considered to be outside the sphere of action of the verb. "Without", "besides", and "but" are typical preposition markers.

All was lost besides this.  
They saved all but the family silver.  
He came without his hat.

Manner (MAN) - The way in which something is done. Manner answers the question 'how?' and typical prepositions are "with" and "without".

He did it without any help.

Possessor (POS) - The individual or thing which has or owns something. The possessor role is not used with inalienable or characteristic possession. It is represented by the possessive (or genitive) case, the preposition "of", or the verb "have".

That book of John's....

Partitive (PR) A portion of something. This role corresponds to what has been referred to as a partitive genitive in many traditional grammars. It is expressed as a genitive in most European languages. In English the construction is only used with "of". The grammatical head noun is frequently a noun of quantity (+MSR) or a part of a larger whole (+PRT).

A piece of pie....  
The chairman of the board....

Respective/Comparative (RES) - A thing or quality being used for comparison. "Than", "with respect to", "as" are typical markers. This is the role which follows comparatives.

John is taller than Mary.

## APPENDIX A-4

### Area of Provenience TAGs

Three different sets of TAGs have been used in the compilation of extensive lexical databases which may be drawn upon at need for the NT system. In the first, values corresponding to fifty fields of activity or areas of study are numerical in form. This older set was devised on the basis of a comparison of the TAGs used in a number of reference dictionaries.

A more extensive set of alphabetical acronyms, derived from the Longman Dictionary of Contemporary English, utilizes from one to four characters for the TAG values. The first two letters indicate some 120 broad fields. This may be extended by another letter following a hyphen, to specify a subordinate field more precisely. Terms broadly relevant in scientific fields, for instance, would receive the value SI (science), while those primarily restrict to physics would be given SI-P. In the accompanying list of alphabetical TAGs, such longer TAGs are listed for medicine (MD), military (MI), and physical science (SI). In the scope notes accompanying the alphabetical tags, asterisks mark the subordinate areas which seem most likely to benefit from extended, 4-character TAGs.

TAGs based on a set of subject codes developed by the German Federal Bureau of Languages have been incorporated as components of entries in a 200,000+ item German/English Transfer Lexicon. These TAGs consist of two letters and two digits. The first letter is a K designating this subject code set; the second signifies some broad field or area of activity. The digits specify subordinate topics within the broad area designated by the second letter.

Sample entries:

(BARB (FIEDER) NST\* (TAG KF57))  
(BARB (STACHEL) NST\* (TAG KF10))  
(BARB (WIDERHAKEN) NST\* (TAG KF30 KZ98))

Numerical TAG Values

01 AEROnautics including aerospace; cf. MILitary

02 AESThetics see also ART, LITERature, METRics, MUSic, THEATre

03 AGRiculture includes animal husbandry; see also BOTany, HORTiculture

04 ANATOMy see also MEDical, PHYSIOlogy, ZOOlogy

05 ARCHitecture cf. MILitary

06 arithmetic see MATHeematics

11 ART includes painting, drawing, etc. See also AESThetics

13 ASTRonomy

athletics see SPORT

15 BIBLical see also RELigion, ECClesiastical

17 BIOlogy see also ANATOMy, PHYSIOlogy, ZOOlogy, BOTany

19 BOTany see also AGRICulture, BIOlogy, HORTiculture

21 BUSiness includes commerce, finance, bookkeeping

23 CHEMistry see also MINeralogy

commerce see BUSiness

25 COMMUnications telegraphy, telephony, radio, video, etc.

27 CULinary

dance see MUSical

29 ECClesiastical refers to organizational and operational aspects of RELigion and its physical appurtenances

31 Electricity includes electronics other than COMMUnications

33 ENGINEering includes mechanics. See also AEROnautics, COMMUnications, ELECTricity, TECHnology

finance see BUSiness

35 GAME largely non-athletic, e.g., bridge, Monopoly, tic-tac-toe; also pick-up-sticks, hop-scotch, billiards; includes gambling, generally

37 GEOGraphy

39 GEOlogy see also MINeralogy

41 HORTiculture see also AGRICulture, BOTany

43 INFormation processing (largely computer-related)

45 LANGuage name

47 LAW

49 LINGuistics exclusive of PHONology or LANGuage name

51 LITERature excluding METRics. See also AESthetics,

53 MARitime includes naval, nautical, etc., but cf. MILitary

55 MATHematics includes arithmetic

mechanics see ENGINEering

57 MEDical includes dentistry. See also ANATomy, BIOlogy, PATHology, PHYSIOlogy, PSYCHology

59 METEORology

61 METRics characteristics of versification, e.g., "iambic"

63 MILitary (use this also to further specify the martial components of naval, AEROnautical, and ARCHitectural terminology)

65 MINeralogy and mining

67 MUSic(al) includes dance. See also AESthetics

mythology see RELigion

naval, nautical see MARitime

69 PATHology see also MEDical

71 PHIlosophy including logic

73 PHONology excluding formal METRics

75	PHOTOgraphy	includes cinematography
77	PHYSics	see also AEROnautics, COMMunications, Electricity, METEORology
79	PHYSIOlogy	see also ANATOMy, BIOLOGY, MEDical, PATHology, ZOology
81	POLitical	includes international and general governmental
83	PUBLISHing	includes journalism
85	PSYCHeology	
87	RELigion	abstract aspects, e.g., theology, mythology. See also BIBIcal, ECClesiastical
89	SCIENTific	see also TECHnology and more specific tags, but this can be used if nothing else fits
91	SOCiology	includes anthropology, ethnology
93	SPORT	athletics, generally, including gymnastics and hunting as well as competitive sports
95	TECHnological	see also ENGINEering, SCIENTific, COMMunications
97	THEATre	see also AESThetics, LITERature, MUSIC
	theology	see RELigion
99	Zoology	see also ANATOMy, BIOLOGY, PHYSIOLOGY

Alphabetic Area of Provenience TAG's

AC\_ architecture --> mouldings, types of building\*

AF\_ art --> fine arts, drawing, painting, art history, techniques  
[Sculpture = SK\_ ]

AG\_ agriculture --> field crops, farm implements, fertilizers,  
pesticides, agribusiness, agrotechny, agrostology,  
horticulture\*, agronomy

AH\_ animal husbandry --> breeds and breeding\*, feeds and feeding,  
housing and equipment, zootechny, bees and  
beekeeping\* /cf. MD-V veterinary medicine/

AL\_ alphabets --> letters of various alphabets (incl. English),  
handwriting [Communications code words = CM\_ ]

AM\_ animal names --> Eng. names used in the systematic classification  
of animals, products derived from  
animals with the same name as the animal,  
e.g., "lamb, chicken" [Words like "genus,  
species" = MD-B biology]

A0\_ aerospace --> rockets\*, missiles, astronautics, satellites,  
capsules, etc. /cf. CM\_ Comsat/

AP\_ anthropology --> physical, ethnobiology\*, ethnography\*, ethno-  
zoology\*, paleontology\*

AR\_ archaeology

AS\_ astromomy

AU\_ automotive --> road vehicles and their parts, motorcycles, hot  
rods, speed shops, etc. [Nonautomotive vehicles  
= VH\_ ; transport regulations = TN\_ ]

BB\_\_ baseball --> softball, stickball, kickball

BD\_\_ building --> bricklaying\*, carpentry\*, house fixtures, masonry, plumbing\*, plastering\*, steamfitting, general contracting

BE\_\_ beauty culture --> cosmetics\*, hairdressing\*, barber stuff, beards, moustaches, sideburns, perfumery\*

BR\_\_ basketball

Bi\_\_ Bible and Apocrypha

BO\_\_ botany (not plant names) --> plant physiology, plant morphology, plant anatomy, plant embryology, plant pathology\* [Systematic names (airae, daisy, fern, etc.) = PM plant names]

BV\_\_ beverages --> brewing\*, bottling, distilling\*, liquors\*, wine\*, cocktails\*, soft drinks\* /cf. FO\_\_ food/

BW\_\_ bowling and bowls --> bowls (lawn)\*, candlepin, 9-pin

BU\_\_ business --> advertising\*, commerce\*, marketing\* [cf. advertising], office practice /cf. EC\_\_ economics/

CA\_\_ calendar --> months, days, Hebrew & Arabic calendars\*, holidays\*, time zones\*, seasons and related adjs. /cf. Nb numbers/

CC\_\_ compass --> points of compass; north, east, south, west, and relatives

CD\_\_ card games --> whist\*, poker\*, bridge\*

CE\_\_ ceramics --> earthenware, pottery, porcelain, tilemaking [Glass = GI\_\_ ]

CC\_\_ cartography --> maps, coast surveying

CK\_\_ cricket (sports)

CL\_\_ clothing --> shoes + shoemaking\*, gloves, hats, hosiery, costumes, furs\*, umbrellas

CO\_\_ communications --> telephones\*, telegraph\*, cables\*, Comsat, communications code words

CO\_\_ color(s)

CS\_\_ closed system part-of-speech categories (i.e., words in non-lexical parts-of-speech-- the determiners, conjunctions, prepositions, auxiliary verbs, pronouns [Words like "pronoun, preposition" = LN\_\_])

CT\_\_ court games --> fives\*, handball\*, paddle ball, squash\*, racquets, shuffleboard, jai alai

DA\_\_ dance --> ballet\*, choreography\*

DF\_\_ dentistry --> orthodontics, etc.

DG\_\_ drugs (not pharmacy) and drug slang --> pot & hemp family, opium, etc., LSD & psychedelia, drug culture and slang /cf. RX\_\_ pharmacy/

DP\_\_ data processing --> computers, machine languages, punch cards, hardware, software, programming, OCR, etc.

EC\_\_ economics --> finance, banking\*, stock exchange\*, accounting\*, bookkeeping\*, investment\*, mortgages\*, taxation\*, real estate\* /cf. BZ\_\_ business/

ED\_\_ education --> schools\*, degrees\*, academic dress, academic subjects, audio-visual aids, PTA

EG\_ engineering --> civil\*, chemical\*, mechanical\*, electronics\*,  
surveying\* [Mining & petroleum engineering =  
MG\_ ]

EQ\_ equestrian --> handling, harness, farriery (no breeds), horse  
riding /cf. SP\_ sports; HK\_ hockey/

ER\_ epithets (abusive) --> racial and religious

FA\_ firearms (not obviously military)

FB\_ football --> US\*, Canadian\*, Australian\*, soccer\*, rugby\*

FD\_ fire department

FO\_ food --> cookery\*, cheeses, dietetics, flour milling, meat packing  
and cutting, spices, TV dinners, food processing\*,  
canning, tea, coffee, meals, baking\*

FR\_ forestry --> forest technology, lumber and lumbering, woods

FU\_ furniture and antiques --> modern and antique, office furn.,  
shelves, antiques generally\*

GA\_ games --> croquet\*, board games, checkers\*, chess\*, darts, quoits,  
horseshoes, pool, billiards\*, bumper pool, bagatelle,  
dominoes [Card games = CD\_] /cf. SP\_ sports/

GB\_ gambling (except cards) --> bookmaking\*, gambling terms used in  
lotteries\*, football pools  
[Card games = CD\_ ]

GF\_ golf

GL\_ glass

GO\_ geology and geography --> geophysics, hydrology, petrography, petrology, physiography, seismology\*, stratigraphy, geochronology, geologic time and eras

GY\_ gymnasium sports --> gymnastics, boxing\*, fencing\*, wrestling\*, weight lifting, judo\* and karate\*

RA\_ handicrafts --> basketry, brickmaking, cabinetwork, cooperage, joinery, locksmithing, plating, taxidermy, tinsmithing, woodturning, do-it-yourself\* handicrafts [Tools, screws, parts = HW\_ hardware]

RE\_ heraldry

HF\_ hunting and fishing --> angling\*, falconry\*, safaris, trap-setting, archery, fox-hunting\*, whaling

HH\_ household --> household equip., interior decorating, dishes, domestic do-it-yourself [Knitting-&-sewing = KS\_]

HI\_ history -- dynasties, historical eras and ages, ancient Rome and Greece, feudal system

HF\_ hockey and other field games specified --> ice\* and field hockey, lacrosse\*, polo\*

HR\_ horology --> timepieces, clocks and watches

HW\_ hardware --> hand tools\*, screws, nails, hinges, machine parts

IN\_ insurance

IS\_ information science --> documentation, information retrieval, indexing (FWIC, coordinate)

JW\_ jewelry --> terms, gems (N.B. conflict with MN\_ mineralogy /cf. MN\_ mineralogy (lapidary work))

KL\_ handweapons (not firearms) --> knives, swords, bludgeons

KN\_ knots --> including ropes and rope-making, cordage

KS\_ knitting\*-and-sewing\*

LA\_ language names {N.B. potential confusion between this and LN}

LB\_ labor --> unions, trade-union legislation and terminology

LN\_ linguistics and grammar --> phonetics\*, semantics\*, modern grammars (generative, transformational\*, stratificational, etc.), spelling, types of sentence [Information theory = MH\_ ; language names = LA]

LT\_ literature --> criticism, lit. forms, lit. history, drama\* (as distinct from theatre), prosody\*

LW\_ law [trade union legislation = LB\_ ; traffic regulations = TN\_ ]

MD-A anatomy [cytology (cells) = MD-Y; histology (tissue) = MD-H]

MD-B biology (evolutionary, natural history, misc.) [Biochemistry = MD-C; microbiology = MD-O; botany = BO\_ ; zoology = ZO\_ ; animal names = AM\_ ; plant names = PM\_ ]

MD-C biochemistry

MD-D pathology (not of plants)

MD-E ecology (plant & animal ecology, biogeography, phytosociology, conservation, limnology)

MD-F chiropody

MD-G genetics & gynaecology

MD-H histology

MD-I immunology {N.B. also listed under physiology}

MD-L obstetrics

MD-M medicine (the field) [Pharmacy = RX\_ ]

MD-N neuroscience and neurology

MD-O microbiology (bacteriology <medical, systematic, industrial>, virology)

MD-P physiology (general, animal & human, ethology, immunology\* {N.B. conflict with MD-I}, serology)

MD-Q osteopathy

MD-R embryology (developmental animal biology)

MD-S surgery

MD-T ophthalmology  
MD-V veterinary medicine  
MD-Y cytology (structure & junction of cells, incl. histochemistry & ultrastructure (electron microscopy))  
MD\_ medicine and biology, generally

MD\_ manufacturing e.g.: brushes, candles, cement, luggage, leather\*, & tanning\*, matches, rubber, soap, linoleum, methods of manufacturing

MD\_ mining engineering --> mines, petroleum & natural gas, quarrying\*

MD\_ mathematics --> usual subdivisions, set theory, games theory, new math, information theory, arithmetic, algebra\*, trigonometry\*, geometry\*

MI-A army [Ranks = MI-R]  
MI-B ballistics  
MI-F fortifications  
MI-G guns (military), ordnance  
MI-K armour  
MI-N navy  
MI-P air force  
MI-R ranks, military  
MI-V naval vessels, names of types  
MI\_ military (general-- not classifiable under army\*, navy\*, or air force\*)

MI\_ meteorology --> climatology

MI\_ mineralogy --> gems {N.B. conflict with JW\_ jewelery, q.v.}, lapidary work, crystallography, minerals

MI\_ motion pictures --> production, projectors, slides  
[Broadcasting = RA\_]

MI\_ measures and weights --> standards, systems, units of measure\*

MI\_ metallurgy --> founding\*, smelting\*, patternmaking, blacksmithing\*

MI\_ music --> bells, change-ringing\*, organs, piano manuf., instruments, singing, opera, jazz\*, pop\*, rock\*, and folk\*

MY\_\_ mythology and legend --> legends, Camelot, etc., legendary beasts

NA\_\_ nautical --> seamanship, navigation, kinds of boats(?) and ships  
[Navy = Mi-N]

NB\_\_ numbers --> numerals, cardinals, ordinals, groups of days, weeks, months, and years

NT\_\_ net games --> tennis\*, table tennis (ping pong\*), badminton\*, deck tennis, volleyball\*

NU\_\_ numismatics --> coins\*, currencies\*, monetary units

OC\_\_ occult --> alchemy\*, astrology\*, palmistry\*, phrenology, spiritualism\*, magic\*, black magic, voodoo

ON\_\_ occupations --> trades

OR\_\_ orders --> awards, medals\*, fraternal orders, Freemasonry\*, knighthood, societies\*, military orders, ranks of nobility or royalty\*, forms of address\*

OZ\_\_ zoology --> ichthyology\* {N.B. conflict: said to incl. fisheries and commercial fishing}, ornithology\*, herpetology  
[Animal names = AN\_\_]

PG\_\_ photography [Motion pictures = MP\_\_]

PH\_\_ philosophy --> ethics\*, logic\*, schools of phil., metaphysics\*

PL\_\_ political science --> civics, diplomatics\*, government, political parties, political economy, Marxism and offspring, slavery, parliamentary procedure  
[Feudal system = HI\_\_]

PM\_\_ plant names --> English names used in the systematic classification of plants, products derived from plants with the same name as the plant [Words like "genus, species" = MD-B biology] /cf. BP\_\_ botany/

PN\_\_ paints --> surface coatings in general, pigments\*

PP\_\_ paper --> stationery, cardboard, paperboard, paper sizes, paper-making

PS\_\_ psychology, etc. --> psychiatry\*, psychoanalysis\*

PT\_\_ printing\* and publishing\* --> bookbinding\*, electrotyping, letter-press, lithography, offset, photo-engraving, typefounding, typography\*, photocomposition, cold composition, journalism\*

RA\_\_ radio --> radio and TV, radar, broadcasting generally  
[Film production, motion pictures = MP\_\_]

RE\_\_ recording --> hi-fi, stereo, tape, recorders, etc.

RL\_\_ religion (Christian\* and/or Jewish\*) --> relig. titles\*, relig. orders\*, theology\*  
[Church architecture = AC\_\_] /cf. BL\_\_  
Bible; RN\_\_ other religions/

RN\_\_ relig N (non-Judeo-Christian, ancient and modern) --> Islam\*,  
Buddhism\*, Hinduism\*, Zoroastrianism\*, Theosophy

RP\_\_ reprography --> xerography, copies, microforms, lithography

RR\_\_ railways --> incl. model railways\*

RX\_\_ pharmacy --> pharmacology, pharmacognosy

SC\_\_ scouting --> Boy Scouts\*, Girl Scouts\*, Girl Guides\*

SI-C chemistry (incl. plastics, chemical equipment) [Elements and  
compounds = SI-E]  
SI-D dyes and dyeing  
SI-E elements and compounds, chemical  
SI-M magnetism  
SI-O optics (light, spectrum, fiber optics, lasers, holography, spectroscopy, microscopy)  
SI-P physics (relativity, fluid mechanics, solid-state, particles, radiography, atomic energy) [Magnetism = SI-M; optics = SI-O; thermodynamics = SI-T]  
SI-T thermodynamics  
SI-X explosives  
SI\_ science (chemical & physical, excluding biological; esp. words common to several fields) /cf. specific sciences/

SK\_ sculpture

SM\_ cruel and unusual punishments --> tortures & torture devices,  
also old-fashioned punishments

SN\_ sounds --> animal sounds\* (moo, chirr), inanimate sounds (clink, clank)

SO\_ sociology --> incl. criminology\*, police, fingerprints, penology, hobos, prostitution, racial problems and terms  
[Racial epithets = ER]

SP\_ sports (catchall for those not in subdivs. BB, BK, BW, CK, CT, FB, GT, GY, HK, NT, TF, WA, WI) --> bullfighting, fieldball, hurling, rodeo, mountaineering\*, camping\*, tenting, etc.

ST\_ stamp collecting --> philately, postal terms\*

TE\_ textiles --> fabrics, rugs

TF\_ track and field --> Olympic events, dashes, hurdles, pole vault, etc., all "man" races, athletics

TH\_ theatre --> stage production terms, acting (the Method, etc.), carnival stuff\*, fairgrounds\*, stage magic, circus\*  
[Drama terms = LI\_]

TN transport --> traffic regulations, freight and shipping, highways, roads

TO tobacco --> pipes, cigars, etc.

VH vehicles (nonautomotive) --> baby carriages, wagons, carriages, scooters, bicycles

WA water sports --> swimming\*, diving, surfing, water polo, scuba diving, water-skiing [Sailing = NA ; rowing = NA ] ]

WI winter sports --> skiing\*, skating\*, bobsleds, skibob, curling\*, sledding, snowmobiles

XX general (but not words in closed system part-of-speech categories, which is CS )

## APPENDIX B

### Grammar Rules

This appendix contains information needed to write grammar rules for the mechanical translation system. The first section describes the format of a grammar rule, and the second section describes the functions that may be used in the different parts of the grammar rule.

#### I. Format

A sample grammar rule:

NC	NST
Ø	1
	(REQ WF)
	(REQ CL A)

TEST

CONSTR	(CPX 1 ALO CL)
	(ADD WF)
	(ADD NU S)
	(ADD CA N G D A)

TRANSF

The first line of a grammar rule consists of the left-hand side (a part of speech) of a rewrite rule, and then the right-hand side parts of speech of the rewrite rule. Each part of speech should be separated by one or more blanks.

The second line consists of the column numbers of the left side (which is always Ø) and the right side (1 to n) parts of speech. For ease of reading, one should align the number with its corresponding part of speech.

Underneath each column number are placed the particular requirements for that given node (column). For each such row, every right-hand side column must have either a requirement test or a "--" which is a place holder. Zero rows (no requirements) are possible. For readability, columns should line up. The application of the grammar rule fails as soon as any test in this section fails. The tests are applied top to bottom, starting with column 1 and working towards the right.

The next part of a rule is for TESTs between constituents. The keyword TEST introduces this part. The tests then follow. The grammar rule will fail as soon as any test fails.

The next part of a rule is for constructing the new father node resulting from the successful application of the rule. The keyword CONSTR introduces this section. The constructors follow.

The last part of a rule is the transfer part. This part is applied in the transfer phase, after the complete syntactic tree is built. The keyword is TRANSF. Functions for transforming a node or its sons and for passing features down into the tree can be performed in this part.

A sequence of ten dashes ('-----') ends each grammar rule.

The following is a more complex example of a grammar rule, with an explanation of each part.

NP	DET	NO	[COMMENTARY]
Ø	1	2	
	(REQ CA)	(REQ CA)	requires: the abbreviations
	(REQ NU)	(REQ NU)	are spelled out below.
	(REQ GD)	(REQ GD)	
	(REQ KD DET)	--	
TEST	(INT 1 CA 2 CA = X1)		intersections; that is,
	(INT 1 NU 2 NU = X2)		agreements in CAse, NUmber,
	(INT 1 GD 2 GD = X3)		and GenDer.
CONSTR	(ADX X1)		With the help of this the noun
	(ADX X2)		phrase is constructed.
	(ADX X3)		
	(ADD PS 3)		
	(CPX 2 CA NU GD)		
TRANSF	(SEF 2 CA)		The result of the intersections
	(SEF 2 NU)		are placed on the NO. The sons
	(SEF 2 GD)		are transferred.
	(XFR)		

-----

This rule builds a noun phrase (NP) from a determiner (DET) and a nominal (NO). In order for the rule to apply, both the DET and the NO must have case (CA), number (NU), and gender (GD) among their features [e.g., as a result of the application of the morphological and word rules]. In addition, the DET must be one whose kind of determiner (KD) is a determiner (DET), since German DETs may also be used as relative pronouns, demonstratives, etc. Assuming these conditions are met, the system then applies the TEST portion of the rule. In this example, the

tests consist of a series of intersection operations which ensure grammatical agreement between the DET and NO with respect to CA, NU, and CD. If the TEST portion succeeds, the CONSTR is applied. The results of these intersections (which are stored in variables X1, X2, and X3, respectively) are then assigned to the newly created NP node. In addition, the NP is assigned the feature person (PS) with value third (3). Finally, all non-system features of the NO node other than CA, NU and GD are carried up to the new node, since they may be needed for later agreement with the verb. Later, when TRANSF is applied to this NP node, the values of X1, X2 and X3 will be sent down (SEF) to the NO node, after which XFR will apply transfer to the DET and NO nodes.

The following list describes the functions that may be used in grammar rules of the translation system. These rules are broken into four categories: requirements on the right-hand side (individual terms); TESTs between constituents; CONSTRs used for building the new node that is the result of a successful rule application; and TRANSF operations.

(: any text) A comment may be used in any one of the four parts of a grammar rule. The comment consists of everything up to the matching right parenthesis, and can include any text, notes, etc...

### 1. Column Tests

Usage	Meaning
(REQ f)	Feature f required to be present (value arbitrary).
(REQ f v1...vn) n >= 1	Feature f required with at least one value in v1...vn
(REQ f * v1...vn) n >= 1	Feature f is required, and must have at least one value not in v1...vn
(NRQ f)	Feature f must not be present
(OPT f v1...vn) n >= 1	If feature f is present, it must have at least one of the values v1...vn
(OPT f * v1...vn) n >= 1	If feature f is present, it must have at least one value not in v1...vn

Note: The system features WI (word initial) and WF (word final) may be tested with REQ and NRQ for placement restrictions.

## 2. TEST

Usage	Meaning
(INT nl f1 n2 f2)	Intersection for features. Succeeds if all elements of feature f1 on term nl are in feature f2 on term n2. Automatically succeeds if f1 is not on nl, or if f1 on nl is empty.
(INT nl f1 n2 f2 = xi)	Intersect value for feature f1 on term nl with feature f2 on term n2; it is assumed both features are present; xi is assigned result of intersection.
(INT xi n f = xj)	Intersect variable xi with feature f from term n; xj is result.
(INT n f v1 ... vj) j>=1	Intersect in if-then-(else) clauses. Intersect the associated value of feature f on the nth son, with the values (v1 ... vj). (Use n=0 for the father node.) Note: v1 may not be a number OR j=2, i.e., it must not look like the form of INT (INT nl f1 n2 f2).
(SUM nl f1 n2 f2 = xi)	Create union of values of feature f1 on term nl and f2 on term n2; assume both present; xi is assigned result.
(SUM xi n f = xj)	Union variable xi with feature f on term n; xj is result; xi can only be result of a previous SUM.
(FRM nl {n2})	Apply the verb case frame to the verb phrase (VP), which is term nl, and to an optional additional argument (e.g. argument that precedes the VP). The case frame fails if the central arguments specified by the transitivity type of the verb can not be found among all the arguments to the predicate according to case, preposition and semantic type. The successful case frame assigns roles to the central arguments of the verb, deletes the preposition in central arguments which are prepositional phrases, and arranges the sons with predicate first, n2 second (if it exists), and then the remaining arguments of the verb phrase.
(FRT nl {n2})	Case frame for use in transformations; identical to FRM except that the new arrangement of sons is placed as the nl-th son, which should be referenced by the new pattern of a transformation.

### 3. CONSTR

Usage	Meaning
(ADD f)	Add feature f to new node (with value T)
(ADD f v1...vn) n >= 1	Add feature f to new node with values v1...vn
(ADF n f)	Add f and its value from term n. Transfer can use n=Ø for transfer of feature f from old father. If f is not present on term n, this is a no-op.
(ADF n f1 f2)	Add f1 from term n, renaming feature to f2. (may use n=Ø for transfer)
(ADX x)	Add variable x's value to new node, using feature name from the original (2nd) argument of the test where x got assigned its value
(ADX x f)	Add variable x's value, using feature name f
(CPX n f1...fi) i >= Ø	Copy all the features and values from term n, except the features in f1...fi or on the internal SYSTEM-FEATURES list.
(CPY n f1...fi) i > Ø	Copy the features (and values of) f1...fi from term n.
(PRF k)	Assign preference factor k to the new node; this factor will be multiplied with the otherwise-calculated weight to determine this node's final weight; if the result falls below CUTOFF-WT, this node will be pruned from further analysis.

Note: All the functions in this section return as their value the new node; they will return NIL if the node is unchanged.

#### 4. TRANSF

(SEF n f {f2}) is like ADF (f2 optional), but adds father's f to nth son; in other words, propagates downwards rather than upwards. SEF returns the nth son.

(SEV n f {vl...vi}) is to ADD as SEF is to ADF; vl...vi default to T. SEV returns the nth son.

(XFM txname)  
(XFM oldpattern newpattern)  
(XFM) Invokes the transformation 'txname' or the transformation defined with 'oldpattern' and 'newpattern'. If neither is specified, the set of general transformations is applied. XFM returns the transformed node if successful, otherwise NIL. (See also Appendix C)

(XFR {i...n}) invokes the function TRANSFER on the sons (of the current father) as indicated by the numbers i...n in that order -- and on no other sons; if i...n is not provided [e.g., (XFR)] then TRANSFER is invoked on all sons. By indicating sons explicitly, it is possible to perform transfer on some sons, raise (via AD\_) and send (via SE\_) features, and then invoke TRANSFER on other (or the same) sons. A terminal node will automatically invoke TLX on that node with no flex. XFR returns the new node.

(XLX n)  
(XLX n (f1...fn))  
(XLX n (f1...fn) (flex f1...fn)) Transfer the nth son of this node with a lexical transfer of its canonical form to an English canonical form, and then create the English lexical entry. Use f1...fn as features to choose the correct sense of the canonical form. If the English node may be suffixed, include the third argument giving the flex name in the dummy lexicon and the features f1...fn to choose the proper sense. The suffixes are concatenated onto the allomorphs of the node. XLX returns the new node.

(ORO) Order the arguments to the predicate using the transitivity type on the predicate, and the roles assigned from the German case frame, apply the English case frame to order the roles. It is assumed the predicate is the first son. ORO returns the new node.

(TLX (f1...fn) (flex f1...fn)) Using the canonical form (CAN) on this node, first transfer this lexical entry using the transfer lexicon. Then using the new CAN, create the lexical entry using the target language lexicon. f1...fn if present, are features used to check agreement of

the old node and the new node. As in XLX, if the second argument is present, create a dummy and concatenate the ALO to the node's ALO. This function is to be used in transformations. It returns the new node.

5. The following functions may be used to perform if-then and if-then-else contructions in a grammar rule. They may be used in any portion of a rule described above. Care must be taken in that the result returned must be a function that could originally be called in these parts of a grammar rule.

OR - this Lisp function used in the TEST part of a grammar rule allows a rule to succeed if one of its tests succeeds.

Form: (OR t1 t2 ... tn)

It succeeds if at least one of tests t1 ... tn succeeds, evaluated left to right.

AND - this Lisp function may be used when one conditionally wants to add a feature or alter preference. It is an if-then construction. The form of the function call is:

(AND t1 t2 t3 ... tn v)

If all of the ti tests succeed (evaluated left to right), then return/perform v.

COND - this Lisp function is a way of expressing if-then-else. If the first part of any pair succeeds, the second is returned as the value. Form of a call:

```
(COND (t1 v1)
      (t2 v2)
      (t3 v3)
      .
      .
      (tn vn) )
```

If t1 is true: return/perform v1. If all tests fail, NIL is returned (i.e., no effect). The test t1 may be used as a test that always succeeds (i.e., T means "else").

**RET** - this function may be used to retrieve the values associated with a specific feature on either the father node, or one of its sons. The form of a call is:

(RET n f)

Retrieve the value of feature f on the nth son. Use n=0 for the father node. The value returned is a list of (zero or more) associated values.

**INT** - form: (INT n f v1 ... vj) For an explanation of INT, see above.

**NOT** - this Lisp function may be used for a test to succeed if conditions for a test are not met. NOT succeeds if its argument is NIL or the empty list. It should only be used within a COND, an OR, or an AND.

(NOT t) t is a test

Examples:

```
(OR (NOT (RET 1 NU)) (INT 1 NU 2 NU = X1))  
(AND (INT 1 MD IMP) (PRF .5))  
(COND ( (INT 1 MD IMP NIL) (PRF .5) )  
      ( T (PRF 2) ) )  
(COND ( (INT 2 VC P) (ADD TN FU) )  
      ( T (ADF 2 TN) ) )  
(AND (NOT (INT 1 MD IMP)) (PRF 2))
```

## APPENDIX C

### Transformations

Transformations are used to manipulate subtrees during analysis when constructing a new father node, or during transfer to convert from SL to TL structure. Transformations may be included directly in a grammar rule, or may be invoked by name in a grammar rule. The transformation package is even more general, allowing any transformation which fits the structure to apply, rather than only applying a specific transformation. General transformations were not used anywhere in the METAL grammar, but could have been.

The function XFM is used in a grammar rule to invoke a transformation. The different forms are:

```
(XFM)           Invoke general transformations
(XFM <txname>)  Invoke the transformation <txname>
(XFM <oldpattern>
    <newpattern>) Invoke the specified transformation
```

To define a named transformation, the form is:

```
(DT <oldpattern> <newpattern> <txname>)
```

Named transformations should be in their own files, with all transformations for each phase in a separate file.

A Backus-Naur Form (BNF) description of <oldpattern> and <newpattern> follows:

```
<oldpattern> ::= <pattern> {the exprs cannot change structure}
<newpattern> ::= <pattern> {the exprs may change structure}
<pattern> ::= ( <father> [ ( <son>* ) [ <expr>* ] ] )
<son> ::= <category>:<integer> | &:<integer> | --:<integer> | <pattern>
<father> ::= <category>:<integer> | &:<integer>
```

Notes:

- 1) & matches any single category node
- 2) -- matches an arbitrary number of sons, including zero
- 3) <integer> is used for referencing nodes either for tests or for equating nodes in <oldpattern> with <newpattern> when ambiguous
- 4) \* indicates zero or more occurrences allowed
- 5) () zero sons, must be used to fill the sons slot if no sons exist or are not needed, and an <expr> is necessary
- 6) Column tests performed in <pattern> refer to <father>
- 7) Test between constituents refer to the <integer>s in the transformation
- 8) <expr> that can not change structure is those in column tests and TEST
- 9) <expr> that changes structure is those used in CONSTR and TRANSF

## APPENDIX D

### Case Frames

This appendix describes the individual case frames in detail for parsing German and generating English. A frame uses voice and mood features to decide which subframes to attempt to apply. The subframes are ordered by trying the one that covers the most arguments first. In coding transitivity type for a verb, choose all possible frames for the verb, including shorter frames where one argument is optional and the longer frame does not allow optionality (e.g., both T2AT and T3ATR if recipient is optional).

#### Notes:

- 1) A case frame will fail to apply if the non-central arguments to the frame are not legal as peripheral arguments. Legal peripheral arguments are any adverb except "nicht", and prepositional phrases whose prepositions can mark one of the following roles: LOC, TMP, DST, ORN, ART, TOP, EQU, MAN, PUR, EXC, INS, OPP, COM, PRT and MSR.
- 2) 'Verb-specified' indicates that the information from the lexical entry of the main verb in the predicate is used to determine marker (grammatical case or specific preposition) and form of argument if not otherwise stated.
- 3) E: <role 1> - PPED - x <role 2> - (x <role 3>) indicates that the English ordering of given subframe for generation is <role 1> before the predicate, <role 2> immediately after the predicate with 'x', which means that the English verb entry is checked for a marker, and if an optional <role 3> exists, place it after <role 2> with verb-specified marker (x). The peripheral arguments are placed after all central arguments in the same sequence as in German.
- 4) The first nominal argument specified in a frame must agree in person and number with the predicate.
- 5) See appendix A-3 for a list of the case role definitions used below.
- 6) A pronoun which is co-referential with another argument must agree in person, number, and gender with that argument. For example: see I2AX.
- 7) 'any' role = no restrictions on the role. Currently the system will only recognize a prepositional phrase (PP) that marks that role. The preposition of this general PP is NOT deleted. (Eventually semantic type of a noun phrase (NP) may also identify this role.) For example: see I2AX.

### I1A

Intransitive verb with one central argument, an agent (AGT)

Active: requires nominative NP as AGT

E: AGT - PRED

Passive: no passive form

Imperative: no arguments required

E: PRED

### I1T

Intransitive verb with one central argument, a target (TAR)

Active: requires nominative NP as TAR

E: TAR - PRED

Passive: no passive form

Imperative: no arguments required

E: PRED

### C2AQ

Copula type verb with two central arguments, an agent (AGT) and an equative (EQU)

Active: requires a nominative NP as AGT, and a verb-specified EQU

E: AGT - PRED - x EQU

Passive: none

Imperative: none

### T2AT

Transitive verb with two central arguments, an agent (AGT) and a target (TAR)

Active:

a) Stative verb - requires nominative NP as TAR, and takes optional verb-specified ACT

E: TAR - PRED - (x AGT)

b) Non-stative

1) requires nominative NP as AGT and verb-specified TAR, with an optional PP marked by "mit" as instrument (INS)

E: AGT - PRED - x TAR - (x INS)

2) or requires nominative NP as INS and a verb-specified TAR

E: INS - PRED - x TAR

Passive:

a) Regular Passive - requires nominative NP as TAR, and takes optional verb-specified PP as AGT, and optional verb-specified PP as INS

E: TAR - PRED - (x AGT) - (x INS)

b) Impersonal passive construction with understood "es" -

Optional verb-specified PP as AGT and verb-specified PP as INS, the pronoun "es" is added as a TAR to the clause

E: TAR - PRED - (x AGT) - (x INS)

Imperative: requires a verb-specified TAR and optional verb-specified PP as INS

E: PRED - x TAR - (x INS)

**T3ATR**

Transitive verb with three central arguments, an agent (AGT), a target (TAR), and a recipient (REC)

Active: requires nominative NP as AGT, verb-specified TAR and verb-specified REC

E: AGT - PRED - x TAR - x REC

Passive: requires nominative NP as TAR, verb-specified REC, with an optional verb-specified PP as AGT

E: TAR - PRED - (x AGT) - x REC

Imperative: requires verb-specified TAR, and verb-specified REC

E: PRED - x TAR - x REC

**T3ATB**

Transitive verb with three central arguments, an agent (AGT), a target (TAR), and a benefactive (BEN)

Active: requires nominative NP as AGT, verb-specified TAR and verb-specified BEN

E: AGT - PRED - x TAR - x BEN

Passive: requires nominative NP as TAR, verb-specified BEN, with an optional verb-specified PP as ACT

E: TAR - PRED - (x AGT) - x BEN

Imperative: requires verb-specified TAR, and verb-specified BEN

E: PRED - x TAR - x BEN

**C2AJ**

Copula type verb with two central arguments, an agent (AGT) and an equitive (EQU) adjective (ADJ)

Active: requires nominative ADJ as EQU and nominative NP as AGT

E: AGT - PRED - EQU

Passive: none

Imperative: requires nominative ADJ as EQU

E: PRED - EQU

**I2AX**

Intransitive verb with two central arguments, an agent (AGT) and a reflexive pronoun (RFX)

If successful, delete RFX, change the TT to I1A, add RFXDEL feature to PRED

Active: requires accusative reflexive pronoun as RFX and nominative NP as AGT, RFX must be co-referential with the AGT

E: AGT - PRED - x RFX

Passive: none

Imperative: requires accusative reflexive pronoun as RFX

E: PRED - x RFX

### I2AL

Intransitive verb with two central arguments, an agent (AGT) and a locative (LOC)

Active: requires nominative NP as AGT, and any LOC  
E: AGT - PRED - LOC

Passive: none

Imperative: requires any LOC  
E: PRED - LOC

### T3ATX

Transitive verb with three central arguments, an agent (AGT), a target (TAR), and a reflexive pronoun (RFX)

If successful, delete RFX, change TT to T2AT, add RFXDEL feature to PRED

Active: requires verb-specified reflexive pronoun as RFX, nominative NP as AGT and verb-specified TAR, RFX must be co-referential with AGT

E: AGT - PRED - x TAR - x RFX

Passive: none

Imperative: requires verb-specified reflexive pronoun as RFX, and verb-specified TAR  
E: PRED - x TAR - x RFX

### T2AR

Transitive verb with two central arguments, an agent (AGT) and a recipient (REC)

Active: requires nominative NP as AGT, and verb-specified PP as REC  
E: AGT - PRED - x REC

Passive: none

Imperative: requires verb-specified PP as REC  
E: PRED - x REC

### C2AM

Copula type verb with two central arguments, an agent (AGT) and a measure (MSR)

Active: requires a nominative NP as AGT, and a verb-specified MSR  
E: AGT - PRED - x MSR

Passive: none

Imperative: none

### C3ATM

Copula type verb with three central arguments, an agent (AGT), a target (TAR), and a measure (MSR)

Active: requires a nominative NP as AGT, a verb-specified TAR and a verb-specified MSR

E: AGT - PRED - x TAR - x MSR

Passive: none

Imperative: none

### M2AC

Modal/auxiliary with two central arguments an agent (AGT) and a complement (CMP)

Active: requires a nominative NP as AGT, and a verb-specified CP as CMP  
E: AGT - PRED - x CMP

Passive: none

Imperative: none

### T3ATC

Transitive verb with three central arguments, an agent (AGT), a target (TAR), and a complement (CMP)

Active: requires nominative NP as AGT, verb-specified TAR and verb-specified CP as CMP  
E: AGT - PRED - x TAR - x CMP

Passive: requires nominative NP as TAR, verb-specified CP as CMP, with an optional verb-specified PP as AGT  
E: TAR - PRED - (x AGT) - x CMP

Imperative: requires verb-specified TAR, and verb-specified CP as CMP  
E: PRED - x TAR - x CMP

### I2AP

Intransitive verb with two central arguments, an agent (AGT) and a purpose (PUR)

Active: requires a nominative NP as AGT and a verb-specified PUR  
E: AGT - PRED - x PUR

Passive: none

Imperative: none

## MISSION of Rome Air Development Center

RADC plans and executes research, development, test and selected acquisition programs in support of Command, Control Communications and Intelligence (C<sup>3</sup>I) activities. Technical and engineering support within areas of technical competence is provided to ESD Program Offices (POs) and other ESD elements. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.

